PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, G06F

A1

(11) International Publication Number: **WO 99/58720**

(43) International Publication Date: 18 November 1999 (18.11.99)

(21) International Application Number: PCT/US99/10387

(22) International Filing Date: 11 May 1999 (11.05.99)

(30) Priority Data:

09/076,668	12 May 1998 (12.05.98)	US
09/292,657	15 April 1999 (15.04.99)	US

(71) Applicant: ACACIA BIOSCIENCES, INC. [US/US]; 12040 115th Avenue N.E., Kirkland, WA 98034 (US).

(72) Inventor: SCHERER, Stewart; 3938 Paseo Grande, Moraga, CA 94556 (US).

(74) Agents: HALEY, James, F., Jr. et al.; Fish & Neave, 1251 Avenue of the Americas, New York, NY 10020 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: QUANTITATIVE METHODS, SYSTEMS AND APPARATUSES FOR GENE EXPRESSION ANALYSIS

(57) Abstract

The present invention provides methods for quantifying the relatedness of a first and second gene expression profile and for ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile. The methods are demonstrated to be useful for quantifying the relatedness of environmental conditions upon a cell, such as the relatedness in effects of pharmaceutical agents upon a cell. The methods are also useful in quantifying the relatedness of a preselected environmental condition to a defined genetic mutation of a cell and for quantifying the relatedness of a plurality of genetic mutations. Also presented are systems and apparatuses for performing the subject methods. Further provided are quantitative methods, systems, and apparatuses for selecting informative subsets of genes for gene expression analysis.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

QUANTITATIVE METHODS, SYSTEMS AND APPARATUSES FOR GENE EXPRESSION ANALYSIS

FIELD OF THE INVENTION

This invention relates to bioinformatic methods applicable to pharmaceutical drug development. More specifically, this invention relates to methods, systems and apparatuses for the quantitative analysis, comparison, storage, and visual display of gene expression profiles. The invention further relates to quantitative methods, systems, and apparatuses for the selection of informative subsets of genes for expression analysis.

BACKGROUND OF THE INVENTION

In traditional drug discovery efforts, a specific drug target, such as an enzyme in a known biochemical pathway, is first selected. Next, one or more *in vitro* or *in vivo* assays specific to the chosen target must be developed. Only after the target is chosen and specific assays developed can chemical compounds be screened for the desired activity. Once

compounds are identified that have the desired activity against the chosen target in the dedicated assays, these initial lead compounds serve as the structural predicates for developing derivatives with more favorable therapeutic, pharmacokinetic, and clinical properties. The bioactivity of these derivatives is often assessed using the same dedicated assays which identified the lead compound.

Each of the above-described steps in the traditional drug development paradigm contributes to the risk that a drug showing promise in preclinical testing will fail in clinical trials.

First, selection of the drug target presupposes knowledge of the biological pathways that are clinically relevant to the disease or pathologic process for which the drug is intended. Once clinical testing begins, the chosen target may prove to be physiologically unsuitable. The target may, for example, be involved in a number of related or unrelated biological pathways. The dedicated *in vitro* assays may fail to identify effects of the candidate drug on these parallel or intersecting biological pathways. As a result, drugs that desirably affect the target's activity *in vitro* may prove unacceptably toxic or present undesirable side effects when administered *in vivo*.

Second, the *in vitro* assay methods may themselves prove to be insufficiently sensitive, insufficiently specific, or both. The use of the same assays in the development of derivatives of the lead compound may compound these problems.

There thus exists a need in the pharmaceutical arts for improved strategies for drug

development. In particular, there exists a need for a drug development scheme that depends less upon the initial selection of a suitable target. There also exists a need for a drug development strategy that avoids the isolation, during preclinical drug development, of the selected target from the biological pathways of which it is a part. There further exists a need for a method of drug development that identifies biological pathways and new targets relevant to the pathologic state, disease or disorder of interest.

Recent technical advances in measuring gene expression have made possible the contemporaneous measurement of the expression of many, if not all, genes transcribed in a prokaryotic or eukaryotic cell. The ability to generate such gene expression profiles offers the raw material from which a new drug development strategy may now be fashioned. See, *e.g.*, Ashby *et al.*, United States Patent No. 5,549,588.

Most gene expression profiles to date have been generated by isolating nucleic acid expression products from a host cell, labeling the products (*e.g.*, with a fluorescent or radionuclide label), and hybridizing the labeled nucleic acids to a spatially-addressable matrix comprising units with surface-immobilized DNA having discrete sequences. See, *e.g.*, Lashkari *et al.*, Proc. Natl. Acad. Sci. USA, 94, pp. 13057-62 (1997); DeRisi *et al.*, Science, 278, pp. 680-86 (1997); Wodicka *et al.*, Nature Biotechnology, 15, pp. 1359-67 (1997); and Pietu *et al.*, Genome Research, 6, pp. 492-503 (1996).

The elements of the matrix are selected to represent the totality of genes that can be expressed by the host from which the immobilized DNA matrix was

prepared. Specific hybridization to various DNA elements in the matrix, as recorded, *e.g.*, by scanning laser, scanning confocal fluorescence microscopy, or PhosphorImager, indicates expression of the respective gene. The identity of the respective gene is encoded in the spatial location of the element in the matrix. The data are acquired, digitized, and stored electronically. Taken together, the data identify the subset of genes expressed by the chosen cell culture.

Ashby *et al.*, U.S. Patent No. 5,549,588 (incorporated herein by reference), disclose an alternative approach to generation of gene expression profiles. Ashby discloses a "genome reporter matrix" in which, in one embodiment, each element of the spatially-addressable matrix consists of one or more identical cells (or clones of cells), rather than of specific nucleic acid sequences. The cells at each matrix position contain a recombinant construct that directs expression of a common reporter gene from a distinct transcriptional regulatory element. The transcriptional regulatory element may be drawn from any number of potential eukaryotic or prokaryotic organisms. A sufficient number of matrix elements, and thus of transcriptional regulatory elements, is included to provide a representative sampling of the gene expression repertoire of the chosen organism.

To measure gene expression, Ashby *et al.* read the matrix directly by scanning with a detection device as appropriate to, and dictated by, the reporter. In one embodiment, the reporter encodes a protein that generates a fluorescent signal, such as green fluorescent protein, and is thus scanned with a fluorescence detector; in another embodiment, the

reporter encodes enzymes that produce signals detectable photometrically, and is scanned with a photometer. Signals, as recorded by the scanner, indicate expression operably controlled by the respective transcriptional regulatory element, the identity of which is encoded in the spatial location of the element in the matrix.

Each of the above-described technologic platforms for generating gene expression profiles, herein collectively termed "expression matrices", generates a large amount of information about the concurrent expression of genes in a cell under defined conditions. Such a gene expression profile, in its totality, captures the global gene expression state of the cell under a chosen set of environmental conditions.

The art has heretofore emphasized qualitative comparisons of such gene expression profiles, such as the identification of the subset of genes that show altered levels of expression under different conditions; alternatively, the art has emphasized data manipulations that are not amenable to the quantitative comparison of large, multidimensional data sets. See, e.g., Ashby *et al.* (*supra*); Lashkari *et al.*(*supra*); DeRisi *et al.*(*supra*); Rine *et al.*, WO 98/06874; and Seilhamer *et al.*, WO 95/20681 (each of which is incorporated herein by reference).

None of these qualitative analysis methods permits the reproducible calculation of relatedness of entire gene expression profiles. It would thus be advantageous to generate quantitative gene expression profiles and, with that information, to compare quantitatively the relatedness of gene expression in a

chosen cell under varying environmental conditions (*e.g.*, treated with different compounds).

Thus, there exists a need for methods that quantify the relatedness of a first and second gene expression profile. There further exists a need for methods that permit the ordered ranking of the relatedness of a plurality of gene expression profiles to a single pre-selected gene expression profile. And there exists further need for quantitative methods and apparatuses that would permit stored data sets (*i.e.*, gene expression profile data from prior experiments) to be queried and analyzed in new comparisons for relatedness.

Although recent technical advances in measuring gene expression have made possible the contemporaneous measurement of the expression of many, if not all, genes transcribed in a prokaryotic or eukaryotic cell, technical considerations often will dictate that fewer than all expressible genes be assayed. For example, samples of drug candidates may be in limiting supply, particularly when produced in small quantity by combinatorial chemistries; there may simply be too little of the agent to permit the testing of its effects on all possible genes of a given cell type. It may also, or in the alternative, be too expensive to assay each candidate agent across each expressible gene of the cell.

These issues are compounded when the genome to be assayed becomes more complex. Thus, to assess the effect of a drug or other environmental agent on each of the expressible genes of a yeast cell, such as *Saccharomyces cerevisiae*, would require the measurement of the expression of about 6,000 genes; to perform the

analogous assay on the gene expression of a nematode, such as *C. elegans*, would require the measurement of the expression of nearly 20,000 genes; to assess the effect of a drug or other environmental agent on each of the expressible genes of a human cell would require the measurement of about 100,000 genes.

Furthermore, not all genes prove equally informative. Some may have an insufficient dynamic range in expression to provide significant information, no matter what the environmental condition. Other genes may vary in expression coordinately, or cooperatively, providing redundancy in the information collected.

One approach to selecting informative subsets of genes for expression analysis is to choose the genes individually by known or suspected function. Thus, Farr *et al.*, U.S. Patent No. 5,811,231 and European patent no. EP 0680517 B1 disclose, *inter alia,* the selection of "stress genes" particularly to identify and characterize compounds that are toxic to the cell.

Such an approach, however, requires antecedent knowledge of the gene's function. Furthermore, the bias imposed by such directed selection would reduce the possibility of identifying previously unsuspected relationships; in a method useful for the identification of such unsuspected relationships, such as the methods presented herein, such directed preselection would be particularly disfavored.

Another approach is to choose the subset entirely at random, in the hope that the subset so selected proves representative of the whole. The problem, clearly, is that the subset so chosen may in

fact prove uninformative for describing the cellular state under one or more environmental conditions.

Yet another approach would be to select genes identified not by common function, but by a common responsiveness to a preselected environmental condition. Whitney *et al.*, Nat. Biotechnol., 16:1329-33 (1998). Falling somewhere between the purely directed and purely random approach, this latter procedure is, to some extent, subject to the disadvantages of both.

There thus exists a need in the art for methods that permit the selection of an informative subset of genes for expression analysis.

SUMMARY OF THE INVENTION

The present invention solves these and other problems in the art by providing methods, systems, and apparatuses for the quantitative analysis of gene expression profiles. The experimental examples demonstrate that such analyses allow one to quantify and to order the relatedness of various drug treatments, permitting the identification of chemical agents that act on the identical molecular target as that affected by a reference drug; permitting the identification of chemical agents that act elsewhere in the same physiologic pathway as that of the reference drug; clarifying the mechanism of action of the reference drug; and clarifying the mechanisms of action of the chemical agents compared to the reference drug--all without the prior identification of the reference drug's molecular target or the development of a dedicated assay. The analyses apply equally to

comparison of other cellular phenotypes, including those caused by other environmental conditions and by genotypic perturbations, including mutations.

In a first aspect, then, the invention provides a method of quantifying the relatedness of a first and second gene expression profile. This first method comprises the steps of: (a) generating a first and second gene expression signal, respectively, for each gene commonly represented in the first and second gene expression profiles; (b) formulating a relative expression score for each pair of first and second gene expression signals; and then (c) calculating from these pair-wise formulated relative expression scores a composite score, the composite score quantifying the relatedness of the two gene expression profiles.

In another aspect, the invention provides a second method of quantifying the relatedness of a first and second gene expression profile, the second method particularly well-suited for comparison of gene expression profiles obtained under mild conditions. This second method comprises the steps of: (a) generating a first and second gene expression signal, respectively, for each gene commonly represented in the first and second gene expression profiles; and then (b) performing a linear regression on the set of paired first and second gene expression signals for the commonly represented genes; wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

In a third aspect, the invention provides a method of ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising the steps of: (a) quantifying pairwise the relatedness of each of the

plurality of gene expression profiles to the preselected gene expression profile; and then (b) ordering said pairwise-measured quantities. In preferred embodiments of this aspect of the invention, the pairwise quantification of relatedness is performed according to one of the two methods newly described herein.

In one series of embodiments of the aforementioned methods, the invention provides methods of quantifying the relatedness of a first and a second environmental condition upon a cell, comprising the steps of: (a) obtaining from the cell, or from genotypically identical cells, a gene expression profile under each of the first and second environmental conditions; and then (b) quantifying the relatedness of the first and second gene expression profile. In preferred embodiments, the first and second environmental conditions each comprises exposure to a chemical compound, such as a pharmaceutical agent.

The invention further provides methods for ordering the relatedness of a plurality of environmental conditions to a single preselected environmental condition upon a cell, comprising the steps of: (a) obtaining from the cell, or from genotypically identical cells, a gene expression profile for each of the plurality of environmental conditions and for the preselected environmental condition; (b) quantifying pairwise the relatedness of each of the plurality of gene expression profiles to the preselected gene expression profile; and then (c) ordering the pairwise-measured quantities. In preferred embodiments, the environmental conditions comprise exposure to a chemical compound.

In another set of embodiments, the invention provides methods of quantifying the relatedness of a preselected environmental condition to a defined genetic mutation of a cell, comprising the steps of: (a) obtaining a first gene expression profile from a cell bearing the defined mutation and a second gene expression profile from a wild-type cell under the preselected environmental condition; and then (b) quantifying the relatedness of said first and second gene expression profile.

The invention further provides methods of ordering the relatedness of each of a plurality of environmental conditions to a defined genetic mutation of a cell, comprising the steps of: (a) obtaining a set of first gene expression profiles from a wild type cell under each one of the plurality of environmental conditions and a second gene expression profile from a cell having the defined mutation; (b) quantifying pairwise the relatedness of each of the first gene expression profiles to the second gene expression profile; and then (c) ordering the pairwise-measured quantities. In preferred embodiments, the environmental conditions comprise exposure to a chemical compound, and the pair-wise quantification is performed according to one of the two methods newly presented herein.

In another series of embodiments, the invention provides methods of quantifying the relatedness of a first genetic mutation of a cell to a second genetic mutation of a cell, comprising the steps of: (a) obtaining a first gene expression profile from a cell having the first genetic mutation and a second gene expression profile from a cell having the second genetic mutation; and (b) quantifying the relatedness

of the first and second gene expression profile. The invention further provides methods of ordering the relatedness of each of a plurality of genetic mutations to a preselected genetic mutation of a cell, comprising the steps of: (a) obtaining a set of first gene expression profiles from cells each having one of the plurality of genetic mutations and a second gene expression profile from a cell having the preselected mutation; (b) quantifying pairwise the relatedness of each of the first gene expression profiles to the second gene expression profile; and (c) ordering the pairwise-measured quantities.

In preferred embodiments, the environmental condition includes exposure of the cell to a chemical compound, the cell is a yeast cell, preferably *Saccharomyces cerevisiae*, and the gene expression profile is acquired from a genome reporter matrix. The methods may broadly be applied, however, to any environmental condition, prokaryotic as well as eukaryotic cells, including human cells, and to gene expression profiles obtained from other types of expression matrices.

In another aspect, the present invention provides systems, including computer systems, for performing the aforementioned quantitative methods.

Thus, in one such aspect, the invention provides a system for quantifying the relatedness of a first and second gene expression profile, comprising: (a) means for generating a first and second gene expression signal, respectively, for each gene commonly represented in the first and second gene expression profiles; (b) means for formulating a relative expression score for each pair of first and second gene expression signals; and (c) means for calculating from

the pair-wise relative expression scores a composite score, the composite score serving to quantify the relatedness of the two gene expression profiles.

In a related aspect, the invention provides a system for quantifying the relatedness of a first and second gene expression profile, comprising: (a) means for generating a first and second gene expression signal, respectively, for each gene commonly represented in the first and second gene expression profiles; (b) means for performing a linear regression on the set of paired first and second gene expression signals for the commonly represented genes; wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

And in yet another related aspect, the invention provides a system for ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising: (a) means for quantifying pairwise the relatedness of each of the plurality of gene expression profiles to the preselected gene expression profile; and (b) means for ordering the pairwise-measured quantities.

The invention also provides computer systems for quantifying the relatedness of a first and second gene expression profile, comprising a processor, such as a digital microprocessor, programmed to:
(a) generate a first and second gene expression signal, respectively, for each gene commonly represented in the first and second gene expression profiles;
(b) formulate a relative expression score for each pair of first and second gene expression signals; and then
(c) calculate from the pair-wise relative expression scores a composite score, wherein the composite score

quantifies the relatedness of the two gene expression profiles.

Analogously, the invention provides computer systems for quantifying the relatedness of a first and second gene expression profile, comprising a processor, such as a digital microprocessor, programmed to: (a) generate a first and second gene expression signal, respectively, for each gene commonly represented in the first and second gene expression profiles; (b) perform a linear regression on the set of paired first and second gene expression signals for the commonly represented genes; wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

The invention additionally provides computer systems for ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising a processor, such as a digital microprocessor, programmed to: (a) quantify pairwise the relatedness of each of the plurality of gene expression profiles to the preselected gene expression profile; and (b) order these pairwise-measured quantities. Also provided are apparatuses comprising a programmable digital computer, with input means and display means, capable of performing the described computational methods on input expression data and reporting the quantitative results on the associated display means.

In yet another aspect, the present invention provides computer readable storage media storing instructions that, when executed by a computer, cause the computer to perform each of the novel methods herein described, including methods for quantifying the relatedness of a first and second gene expression

profile and methods for ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile.

In a further aspect, the invention provides computer readable storage media containing data structures adapted for the methods of the present invention. In one such aspect, the invention provides a computer readable storage medium containing a data structure configured to store data that quantitatively relate a first and second gene expression profile, the data structure comprising an identifier for each of the expression profiles and a scalar, the scalar quantitatively relating the first to the second gene expression profile. The invention further provides computer readable storage media containing a data structure configured to store data that orders the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising: (a) an ordered list of scalars, each scalar quantifying pairwise the relatedness of each of the plurality of gene expression profiles to the preselected gene expression profile; and
(b) identifiers that associate each scalar with its respective gene expression profile.

Although recent technical advances in measuring gene expression have made possible the contemporaneous measurement of the expression of many, if not all, genes transcribed in a prokaryotic or eukaryotic cell, technical considerations often will dictate that fewer than all expressible genes be assayed. For example, samples of drug candidates may be in limiting supply, particularly when produced in small quantity by combinatorial chemistries; there may simply be too little of the agent to permit the testing

of its effects on all possible genes of a given cell type. It may also, or in the alternative, be too expensive to assay each candidate agent across each expressible gene of the cell.

Thus, in another aspect, the present invention provides methods for selecting informative subsets of genes for expression analysis. The invention provides methods of cellular phenotyping, comprising selecting no more than 20% of a cell's expressible genes for expression analysis, wherein the concurrent expression of the selected genes sufficiently defines the cell's phenotype as to permit the cell's phenotype quantitatively to be related to the phenotype of another cell. In these methods, preferably no more than about 20% of the cell's potentially expressible genes are selected, more preferably no more than about 15% of the cell's potentially expressible genes, even more preferably no more than about 10% of the cell's potentially expressible genes, optimally no more than about 5% of the cell's potentially expressible genes, and in the most preferred embodiments, about 1% - 5%, and even 1 - 2% of the cell's potentially expressible genes. Algorithms for effecting such selection, and computers, systems, networks, and other devices for effecting the methods are also presented.

In one embodiment, the methods of this aspect of the invention comprises selecting, from each group of genes whose expression is correlated, the gene with greatest expressive range. In preferred embodiments, the selection is made from the set of genes commonly represented in a plurality of gene expression profiles, and each of the ranges and each of the correlations is

calculated from expression data in the plurality of gene expression profiles.

In a related aspect, the invention provides a system for selecting an informative subset of genes for expression analysis, comprising: means for selecting, from each group of genes whose expression is correlated, the gene with greatest expressive range. In preferred embodiments, the selection is made from the set of genes commonly represented in a plurality of gene expression profiles, and each of the ranges and each of the correlations is calculated from expression data in the plurality of gene expression profiles.

The invention also provides a computer system for selecting an informative subset of genes for expression analysis, comprising a processor, such as a digital microprocessor, programmed to select, from each group of genes whose expression is correlated, the gene with greatest expressive range; a computer readable storage medium storing instructions that, when executed by a computer, cause the computer to perform a method of selecting an informative subset of genes for expression analysis, the method comprising selecting, from each group of genes whose expression is correlated, the gene with greatest expressive range; and a computer readable storage medium containing a data structure configured to store data that identifies an informative subset of genes for expression analysis, the data structure comprising a set of gene identifiers, optionally including a description of gene function.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other objects and advantages of the present invention will be apparent upon consideration of the following detailed description taken in conjunction with the accompanying drawings, in which:

FIG. 1 is a flow chart describing the process by which gene expression signals suitable for the quantitative analysis of gene expression profiles are derived from signals initially acquired from a gene expression matrix, with **FIG. 1A** schematizing initial signal processing and **FIG. 1B** describing an optional subsequent correction according to an environmentally-matched control;

FIG. 2 is a scatter plot of gene expression signals, as processed according to FIG. 1, derived from genome reporter matrices treated individually with one of two chemotherapeutic agents known to be closely related in structure and function: 50 µg/ml daunarubicin and 50 µg/ml doxorubicin (see Example 2);

FIG. 3 plots gene expression signals derived from matrices treated individually with one of two drugs of disparate structure and disparate function: 50 µg/ml doxorubicin and 0.08 µg/ml miconazole;

FIG. 4 plots gene expression signals derived from matrices treated individually with one of two drugs of disparate structure but similar function: 9 µg/ml mycophenolic acid and 50 µg/ml daunarubicin;

FIG. 5 is a flow chart describing a first process for reducing sets of individual gene expression signals, prepared according to the process schematized in FIG. 1, to values that may be used quantitatively to rank the relatedness of gene expression profiles;

FIG. 6 is a flow chart describing a second process for reducing sets of individual gene expression signals, prepared according to the process schematized in FIG. 1, to values that may be used quantitatively to rank the relatedness of gene expression profiles;

FIG. 7 is a scatter plot of gene expression signals as processed substantially according to FIG. 1, derived from genome reporter matrices comprising 1532 separate gene expression reporters, each matrix treated individually with one of two agents known to be closely related in structure and function: 10 μg/ml Lovastatin (X axis) and 20 μg/ml Mevastatin (Y axis);

FIG. 8 is a scatter plot of gene expression signals from a 96 gene subset of the 1532 gene expression signals presented in FIG. 7, the subset selected according to the algorithm charted in FIGS. 9 and 10;

FIG. 9 is a flow chart schematizing the first of two major steps in an algorithm for selecting informative subsets of genes for quantitative analysis of gene expression profiles; and

FIG. 10 schematizes two full iterations of the second of two major steps in an algorithm for

selecting informative subsets of genes for quantitative analysis of gene expression profiles.

DETAILED DESCRIPTION OF THE INVENTION

In order that the invention herein described may be fully understood, the following detailed description is set forth. In the description, the following terms are employed:

As used herein, the phrase "gene expression matrix" refers to a device for acquiring data on the concurrent expression of a plurality of genes, such as is described in Lashkari *et al.*, Proc. Natl. Acad. Sci. USA, 94, pp. 13057-62 (1997); DeRisi *et al.*, Science, 278, pp. 680-86 (1997); Wodicka *et al.*, Nature Biotechnology, 15, pp. 1359-67 (1997); Pietu *et al.*, Genome Research, 6, pp. 492-503 (1996); Ashby *et al.*, U.S. Patent No. 5,549,588. "Genome reporter matrix" particularly refers to the gene expression matrices of Ashby *et al.*

The phrase "gene expression profile" refers to a data set, however constructed, whether stored permanently or ephemerally, in an electronic medium or otherwise, each element of which set represents a measure of the concurrent expression of a distinct and identifiable open reading frame of a cell, typically as acquired from a gene expression matrix.

In a first aspect, the present invention provides a method of quantifying the relatedness of a first and second gene expression profile, comprising the steps of: (a) generating, for each gene commonly

represented in the first and second gene expression profiles, a first and a second gene expression signal, respectively; (b) formulating a relative expression score for each pair of first and second gene expression signals; and then (c) calculating, from the pair-wise relative expression scores, a composite score, the composite score quantifying the relatedness of the two gene expression profiles.

A second method of quantifying the relatedness of a first and second gene expression profile is also provided, comprising: (a) generating, for each gene commonly represented in the first and second gene expression profiles, a first and a second gene expression signal; and then performing a linear regression on the set of paired first and second gene expression signals for the commonly represented genes; wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

The present invention further provides a method of ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising the steps of:
(a) quantifying pairwise the relatedness of each of the plurality of gene expression profiles to the preselected gene expression profile, using either of the two described methods, and then (b) ordering these pairwise-measured quantities. In preferred embodiments of this aspect of the invention, the pairwise quantification of relatedness is performed according to one of the two methods newly described herein.

Each of these methods may be better understood through reference to the figures, as will now be described in further detail.

Generation of Individual Gene Expression Signals from Initial Expression Data

FIG. 1 is a flow chart describing the process by which gene expression signals suitable for the quantitative analysis of gene expression profiles are derived from signals initially acquired from a gene expression matrix, with FIG. 1A schematizing initial signal processing and FIG. 1B describing an optional subsequent correction according to an environmentally-matched control.

The initial data acquisition steps, delimited by box **116**, may be performed serially, as shown, or may be performed concurrently; digitization **101** may be performed by the signal acquisition device itself, by a separate analog-to-digital converter, or may be obviated by acquiring expression data directly in digital form.

Each of the subsequent data manipulation steps (including those in FIGS. 1A, 1B, 5 and 6) may be accomplished in a programmable digital computer using techniques well known in the computer science art. Some of the steps may alternatively be accomplished using analog circuitry well known in the art. The steps may be performed in a single computing device, a series of computing devices, or distributed in parallel across multiple computing devices, as long as the temporal order of steps is observed. The process may be carried out continuously, as shown, or discontinuously, with intermediate values stored, for example, at the identified steps for subsequent processing.

With respect to the programming of the digital computer, the steps shown in FIGS. 1, 5, 6, 9

and 10 may be coded in any of the higher level languages well known in the art, including but not limited to FORTRAN, BASIC, Pascal, C, C+, C++, Java™, or the like; the results shown in the Figures and presented in the Examples herein were generated using digital computers programmed in C. Alternatively, the steps shown in FIGS. 1, 5, 6, 9 and 10 may be coded directly in assembly language. Many of the steps may also be accomplished using subroutines, macros, or other objects included in commercially available statistical analysis programs such as JMP® (SAS Institute) or UNISTAT® Statistical Package (Unistat, Ltd.) or in programs with mathematical functions, such as Mathematica™ (Wolfram Research, Inc.). The choice of programming language is one to be made by and is well within the skill of the artisan.

As shown in FIG. 1, expression data are first acquired **100** as initial expression signals of a form and in a manner appropriate to the particular gene expression matrix; for the expression matrix of Ashby *et al.*, for example, fluorescence data may be acquired by a scanning laser. The initial expression signals are acquired individually for each of the physical locations of the expression matrix, also termed matrix elements. These initial expression signals represent the level of expression of each of the genes individually assayed in the matrix under a selected environmental condition.

Initial background signals will typically also be acquired, most often concurrently, from one or more control locations on the gene expression matrix. The nature of such background controls depends on the nature of the physical matrix. For example, those matrices that measure hybridization of fluorescently

labeled or radiolabeled nucleic acids may include, as such a control, a measurement from one or more locations on the matrix that contain no nucleic acid at all, or one or more locations on the matrix containing nucleic acid that is not complementary to a known ORF, or both. Analogously, matrices that measure expression from recombinant reporters within transformed cells (see e.g., Ashby *et al.*, *supra*) may include as such a control measurements from one or more locations on the matrix that contain cells lacking the recombinant reporter construct; that contain cells including a recombinant construct unable to express the reporter gene; that contain cells comprising a reporter construct but lacking a necessary substrate; or the like.

Although background control elements will typically be included on each matrix, background measurements may also be acquired from distinct physical matrices, or even historically by reference to earlier stored data values from similar matrices. The choice of the type and number of such controls is well within the competence of the skilled artisan.

The initial expression signals and initial background signals, typically acquired as analog signals representing, for example, intensity of fluorescence, are then digitized **101** and stored electronically as initial signal values and initial background values, respectively. Any convenient tabular, matrix, or spreadsheet format may be used to store these data, which are collectively referred to as a gene expression profile. The data may be stored as volatile data, such as values in random access memory. Alternatively, the data may be stored more permanently

on magnetic, optical, or magnetooptical storage media, or the like.

It will be appreciated that the initial signal value for each distinct element of the expression matrix is separately and distinctly identified, whether by its location in a corresponding multidimensional data matrix, by the appending of header information to each component of the data itself, or by other suitable means known to those skilled in the art. For example, the fluorescence intensity of a single physical matrix element may be represented by a single record with multiple fields, one or more fields of which identify the physical origin of the signal, the date and time of data acquisition, an identifier for the experiment run, and/or the like.

It will further be appreciated that the dynamic range of the initial expression signals will be established by physical limits imposed by the format of the expression matrix, and in particular by the dynamic range of the expression reporter and the sensitivity range of the acquisition device. It will further be understood that the analog signal may be represented as an initial signal value by digital data of varying depths, such as 8-bit, 16-bit, 32-bit, and the like, and that the greater the data depth the finer the distinction in intensity which may be encoded, but the greater the storage requirements for those data. The choice of data depth will therefore be made based upon empiric requirements which will be well understood by the skilled artisan. It will further be understood that initial digitization may be performed using one data depth, with subsequent analysis proceeding with data of

lesser depth. In the latter case, a simple linear transformation may be used to reduce the data depth.

In preferred approaches, floating point numbers are used.

Because the initial expression signals from many matrix locations may be low (*i.e.*, at or even below background) a background correction **118** may preferably, but need not necessarily, be performed. Several methods for making such corrections are known in the art. In one approach, the measured (or historical) background value is added to each initial signal value, irrespective of the input's original value. In another, one-half the measured background value is added to each input value.

Although each of these known approaches or other suitable approaches may be used, the following approach is preferred. Each initial signal value is compared **102** to the initial background value. If the signal value equals or exceeds the background value, no correction is made and the variable Signal is assigned **106** the initial signal value. Alternatively, if the initial signal value is less than the background value, Signal is assigned the value of background **104**.

This preferred approach is more conservative than either of the prior approaches to background correction. Assume a first signal value, A, of zero, and a second signal value, B, equal to the background value (BKG). In the first approach in which BKG is added to each signal value, the value of A becomes equal to BKG, the value of B becomes twice BKG, and B is thus set artificially to twice the value of A. In the second approach in which one-half BKG is added to each signal value, A becomes one-half BKG, B becomes

one and one-half times BKG, and the value of B is thus set artificially as three times the value of A. In the preferred approach, A alone is adjusted to BKG, B remains at BKG, and the value of B after correction is thus no greater than A.

Using such a conservative approach to background correction furthers the goal of the present invention of using as many of the acquired gene expression signals as possible to generate a composite score that relates quantitatively one or more gene expression profiles.

Prior methods have typically assessed changes in cellular gene expression by reporting changes in levels of expression on a gene-by-gene basis. Even when many such genes have been measured contemporaneously, the changes have been reported as a multidimensional data set. *See, e.g.*, Lashkari *et al.*, *supra*. In looking at changes in expression of any one gene, however — or even in looking at changes in the expression of a set of individual genes — the existence of measurement error precludes the use in such comparison of expression signals that change by very little.

It has not been atypical, for example, to disregard those changes in expression level that fail to exceed some chosen multiple of the standard error. For example, it is not atypical to disregard changes in expression of an individual gene of less than two-fold, or less than five-fold, or even of less than ten-fold.

The present invention recognizes, however, that many such disregarded data nonetheless report real changes in gene expression, and can thus contribute useful information to the comparison of gene expression profiles. For example, FIGS. 2, 3, and 4 are scatter

plots, each plotted point of which reports the relative expression of a distinct gene under the two identified conditions. The figures are further described below. For present purposes, what should be appreciated is that the scale is logarithmic, with tick marks on the horizontal and vertical axes of these figures set at intervals of one natural log (e^1, e^2, e^3, etc.). As can readily be seen, much of the data lies within the square delimited by the first tick mark in each direction on the two axes. That is, all of the data within such square would be discarded from the analysis were changes of less than one natural log (approximately 2.7-fold) disregarded for the inability to distinguish such change from the standard measurement error. Were changes of less than two natural logs (e^2, or 7.4-fold) disregarded, all data within the square delimited by the second tick mark in each direction would be eliminated from the analysis. As made evident by the Figures, most of the useful data would consequently be lost.

The present invention makes possible the use of these data. Although the significance of small changes in expression of any one gene may be undeterminable due to the size of the standard error, the significance of the collection of changes may indeed often be determined; where prior methods focus upon the standard error as a measure of significance, the present invention instead focuses upon the standard error of the mean. On average, the collection of changes in gene expression as between two different environmental conditions may be strongly correlated, as further shown below.

Thus, in order to retain as many of the data as possible through the background correction steps,

141, it is preferred to use a conservative correction for background, as set forth above.

The Signal for each matrix element, preferably adjusted for background, is then normalized **108** to control for variance as between otherwise identical experiments, that is, as between data acquisition runs on a single expression matrix, or as between individual data acquisitions from duplicate matrices.

The utility of normalizing expression signals was recognized in the art well before the advances which made possible highly parallel measurements of gene expression using gene expression matrices. Thus, individual gene expression measurements, for example, by Northern blot analysis, were frequently normalized by comparing expression to that of a constitutive housekeeping gene, such as actin, probed either concurrently or serially on the same blot. In this way, variability introduced by unequal gel loading, variation in mRNA purity, or the like, could be controlled.

The limitation of the prior approach was the possibility that the individual gene chosen as the reference standard might itself vary in expression. The problem is compounded in the present invention by the desire to measure the entirety of the gene expression of a cell, including that of "housekeeping genes," and by the desire to measure changes in gene expression in the presence of drugs, the effects of which cannot be predicted *a priori*.

Several methods for normalizing signals to control for variability among experiments exist. One approach assumes the median signal across all genes to

be constant, another normalizes to the root mean square of the signal, and another to the mean log of the signal values. The latter method, normalization to the mean log, effectively damps down outliers, which are those signals furthest in magnitude from the mean signal value.

The preferred method herein is to assume the mean signal, across all genes, to be constant: normalization is thus achieved by dividing each signal by the sum of all signals, as shown **108** (FIG. 1A).

The assumption that the mean gene expression signal should be constant may not be valid, however, when only a small percentage of a cell's expressed genes is assessed. Thus, when a small subset of genes is chosen — for initial generation of gene expression profiles, for subsequent quantitative analysis, or for both initial acquisition and subsequent analysis — the normalization step may optionally be omitted. Accordingly, normalization step **108** was omitted from the analysis of the 96 gene subset during the quantitative analyses reported in Example 5, below: the normalization step was omitted because the assumption of constant mean expression may not prove valid.

As a final step **110** in preparing the individual signal values for quantitative gene expression profile analysis, the logarithm of each signal value is taken; that is, Signal is assigned the logarithm of the Signal value. The natural logarithm is preferred, although $\log_{10}$ may also be used.

There are three advantages to performing the comparative analysis using the logarithm of the signal value. First, conversion to logarithmic values allows

equivalent fold-changes in expression levels to be assessed equivalently, whether such change is an increase or decrease in expression.

Consider, for example, a ten-fold decrease and a ten-fold increase from an initial value of one. The 10-fold decrease, to 0.1 units, is an absolute decrease of 0.9 units. A ten-fold increase, to 10 units, is an absolute increase of 9.0 units. The absolute value of the increase, 9.0, appears a far greater change in gene expression than the absolute 10-fold decrease of 0.9 units. Taking the $\log_{10}$ of each value, in contrast, gives values of -1, 0, and +1 for the three values, and the increase and decrease appear identically significant.

An additional although ancillary advantage of calculating logarithmic values is that the quality of the expression profile data set may be directly assessed. The log ratios calculated for all genes, when two replicate profiles are compared, appear to distribute about zero according to a Normal distribution from random measurement errors. Standard statistical measures thus permit quantitation of the degree of reproducibility of the measurements as between different experiments.

The third advantage of using logarithmic values is that plotting the values on a logarithmic scale presents advantages in the visual display of data, as demonstrated in FIGS. 2 - 4 (see below).

The Signal that results from the process of FIG. 1A, concluding with step **110**, is suitable for use in the quantitative analysis of gene expression profiles, as further schematized in FIGS. 5 and 6. However, a series of additional steps, as set forth in FIG. 1B, is preferably performed.

Drugs are formulated in various solvents, including organic solvents, which themselves may variously affect gene expression. Thus, changes in a gene expression profile that result from introduction of a drug into a cell's culture media include changes (1) wrought by the drug, and (2) changes caused by the solvent. The media itself may contribute changes, as demonstrated in Example 4 and Table 7, *infra*. Furthermore, strain or cell-type differences may exist as between the cells assayed.

In order to control for these environmental effects, and thus to focus the subsequent profile comparisons solely on the changes in gene expression attributable to the action of the tested drug, the signal from a solvent-matched, media-matched, and preferably strain-matched control should be subtracted, as detailed in FIG. 1B.

First, initial expression signals and initial background signals from a matched control expression matrix are acquired. For example, as a control for the effects on the gene expression profile caused by the presence of methanol in a solution of actinomycin D (Tables 1 and 2, *infra*), an otherwise identical expression matrix (such as a genome reporter matrix), would be treated with methanol alone at the identical concentration, and initial expression signals and initial background signals acquired therefrom.

The correction for the environmentally-matched control is then performed individually for each gene as set forth in FIG. 1B.

First, the gene's Signal from the matched control matrix ($Signal_{mc}$ **132**) is subtracted **134** from the

gene's Signal **130** as acquired from the experimental matrix.

Next, an artifact introduced by the earlier background correction **118**, as followed by normalization, must be addressed by means of two decisional queries, **136** and **140**. The queries may be done sequentially in any order, or may more typically be accomplished in a single line of code.

When the corrected Signal **134** is less than zero -- that is, when $Signal_{mc}$ **132** exceeds the experimental Signal **130** -- there exists the possibility that $Signal_{mc}$ had been artificially and artefactually increased during background correction **104**, as followed by normalization, and that the true value of $Signal_{mc}$ is in fact less than or equal to Signal **130**. Thus, the first decisional query **136** asks whether the corrected Signal **134** is less than zero and if $Signal_{mc}$ was less than its background at step **102**. If the first decisional query **136** returns true, the corrected Signal is set to zero, **138**. That is, because it is impossible to determine whether the corrected Signal is real, the value is set to zero so that the Signal is discarded from subsequent analysis.

Similarly, if the corrected Signal **134** is greater than zero -- that is, when the experimental Signal **130** exceeds the matched control $Signal_{mc}$ **132** -- there exists the possibility that the experimental Signal **130** had been artificially and artefactually increased **104** during background correction, as followed by normalization, and that the true value of Signal **130** is in fact less than or equal to $Signal_{mc}$. Thus, if the second decisional query **140** returns true, the corrected Signal is set to zero **142**.

FIGS. 2, 3, and 4 show scatter plots of gene expression data processed as described above, including the steps set forth in FIG. 1A and FIG. 1B.

The data in FIGS. 2 - 4 are derived from initial expression signals generated by genome reporter matrices (for details, see the Examples below). FIG. 2 plots data derived from matrices treated individually with one of two chemotherapeutic agents known to be closely related in structure and function: daunarubicin and doxorubicin. FIG. 3 plots data derived from matrices treated individually with one of two drugs of disparate structure and disparate function: doxorubicin, a chemotherapeutic agent, and miconazole, an antifungal agent. FIG. 4 plots data derived from matrices treated individually with one of two drugs of disparate structure but related function, mycophenolic acid and daunarubicin, both of which inhibit DNA synthesis.

Each point plotted on the graphs of FIGS. 2, 3, and 4 represents the expression of a specific gene: the X coordinate plots the value as calculated from the signal obtained in the presence of one of the drugs (doxorubicin in FIG. 2, doxorubicin in FIG. 3, daunarubicin in FIG. 4), and the Y coordinate plots the value as calculated from the signal obtained in the presence of the second of the drugs (daunarubicin in FIG. 2, miconazole in FIG. 3, and mycophenolic acid in FIG. 4).

Visual inspection of FIGS. 2, 3, and 4 demonstrates the usefulness of expression profile analysis for facilitating drug discovery, and further demonstrates that at the extremes of relatedness (unrelatedness) presented in these figures, even casual

qualitative analysis of data processed as presented above proves useful.

In FIG. 2, for example, it is readily apparent from casual inspection that the two drugs affect the expression of most yeast genes similarly, if not identically: each gene whose expression is increased by daunarubicin is equivalently augmented by doxorubicin; each gene whose expression is decreased by treatment with daunarubicin is equivalently repressed by treatment with doxorubicin; and those genes whose expression is unaffected by treatment with daunarubicin are similarly unaffected by doxorubicin. The result is that most of the data points lie on a line through the origin.

In contrast, similarly plotted data from gene expression profiles generated using the unrelated drugs doxorubicin and miconazole produce a far different pattern (FIG. 3). As shown in FIG. 3, the expression of some genes is increased by both drugs (those points in the upper right quadrant), the expression of some genes is decreased by treatment with both of the drugs (those points in the lower left quadrant), and the expression of other genes is oppositely affected by the drugs (those points in the upper left and lower right quadrants).

FIG. 4 presents an intermediate case, in which both drugs are known to affect DNA synthesis, albeit by different mechanisms.

Thus, a qualitative assessment of drug relatedness becomes possible. Those drugs (or other environmental conditions) that produce a scatter plot distribution similar to that shown in FIG. 2 are closely related in action; those that produce a distribution similar to that shown in FIG. 3 are

unrelated in action; and those that produce a distribution similar to that shown in FIG. 4 have dissimilar, albeit somewhat related, mechanisms of action.

Given a lead compound of known efficacy, then, it becomes possible to screen derivatives and analogs to identify those with similar activity, without reliance upon a dedicated biochemical assay. In fact, the mechanism of action of the lead compound need not even be known. The potential for such analysis is limited, however, by the ability to recognize such patterns of relatedness. The problem, minimal at the extremes shown in FIGS. 2 and 3, becomes more apparent in the intermediate cases, such as that presented in FIG. 4. This invention addresses this problem by providing a reproducible, quantitative assessment of relatedness of gene expression profiles; the invention additionally permits analysis of greater than two compounds, allowing a ranked order of gene expression profile relatedness to be generated.

Method for Quantifying the Relatedness of Gene Expression Profiles By Generation Of A Composite Score

The present invention provides a method of quantifying the relatedness of a first and second gene expression profile, comprising the steps of: (a) generating, for each gene commonly represented in the first and second gene expression profiles, a first and a second gene expression signal; (b) formulating a relative expression score for each pair of said first and second gene expression signals; and then (c) calculating, from said pair-wise relative expression scores, a composite score, wherein said

composite score quantifies the relatedness of the two gene expression profiles.

The first step of this method has been described above, with reference to FIGS. 1A and 1B. The second and third step are described here with reference to FIG. 5.

In outline, a relative expression score **524** is formulated **528** separately for each gene commonly represented in the two gene expression profiles. Thereafter, a composite score is calculated **526** from the collection of all such individual gene relative expression scores, the composite score serving to quantify the relatedness of the two gene expression profiles.

As detailed in FIG. 5, the signal for a gene under a first condition, Signal1, **500**, is input. This signal has been processed as set forth in FIG. 1; as noted above, the signal has preferentially, but need not have been, corrected as set forth in FIG. 1B by subtraction of an environmentally-matched control. The signal for the same gene under a second condition, Signal2, **502**, similarly processed as set forth in FIG. 1, is subtracted to provide a relative expression score, **504**. Since the signal values input are logarithmic values, **110**, the difference represents a ratio of expression.

An artifact introduced by the earlier background correction **118**, as followed by normalization, must, however, be addressed at this point, as was described above after subtraction of a matched control signal.

The artifact correction is performed using two decisional queries, **506** and **510**. The queries may

be done sequentially in any order, or may more typically be accomplished in a single line of code.

When the relative expression score, Score **504**, is less than zero -- that is, when Signal2 exceeds Signal1 -- there exists the possibility that Signal2 had been artificially and artefactually increased **104** during background correction, as followed by normalization, and that the true value of Signal2 is less than or equal to Signal1. Thus, the first decisional query **506** whether the relative expression score **504** is less than zero and if Signal2 was less than its background at step **102**. If the first decisional query **506** returns true, the relative expression score is set to zero, **508**. That is, because it is impossible to determine whether the relative score is real, the value is set to zero so that the score does not contribute to the composite score **526**.

Similarly, if the relative expression score **504** is greater than zero -- that is, when Signal1 exceeds Signal2 -- there exists the possibility that Signal1 was artificially and artefactually increased **104** during background correction, as followed by normalization, and that the true value of Signal1 is less than or equal to Signal2. Thus, if the second query **510** returns true, the relative expression score is also set to zero **518** so that this relative score does not contribute to the composite score.

Next, a gene-by-gene threshold comparison is made, **522**. Each expression matrix technology has its own detection threshold below which signals cannot reliably be measured. For example, the oligonucleotide hybridization platform of Lashkari *et al.*, *supra*, has a

different detection threshold from the cellular genome reporter matrix of Ashby *et al., supra.*

Such thresholds are determined empirically. In a simple approach, one twice performs the identical experiment, whether acquisition of a no-treatment profile, or acquisition of a profile from cells identically treated by the same drug. The log ratios calculated for all genes, when two replicate profiles are compared, appear to distribute about zero according to a Normal distribution (provided there is a reasonable signal-to-noise ratio-- if the signals are low, the background correction distorts the distribution) due to random measurement errors. The standard deviation of this distribution provides a guide for setting an appropriate threshold.

Thus, if the absolute value of the relative expression score, as corrected **514** for background artifact, is less than an empirically set threshold, **516**, Score is assigned a value of zero, **518**, and will not thereafter contribute to the composite score, **526**. At present, the preferred threshold for data acquired from the genome reporter matrix of Ashby *et al.* is 0.7. The skilled artisan will be able to establish such empirical thresholds using the statistical techniques above-described. Furthermore, as technology changes and/or those acquiring data become more proficient with existing data acquisition technologies, this empirical threshold will likely change. In experimental Examples 1 - 4 that follow herein, using data earlier collected, a threshold of 1.0 was applied.

It should be noted, too, that the steps delimited by box **522** also remove from further consideration the direction of the change in the

expression of a gene as between a first and second gene expression profile. This is of course necessarily the case for Scores set to zero **518** for failure to exceed the user-defined threshold. As for the remaining scores, the directionality is eliminated by the assignment of the absolute value **520** of any non-negative scores to Score. In measuring the relatedness of two treatments, the informational content of a gene's repression is thus treated equivalently to that of a gene's activation-- only the magnitude of the relative change is used.

Thus, it can be seen that there are two steps in the algorithm at which the relative expression score is set to zero and the data thus eliminated from contributing to the composite expression profile score. In steps **506**, **508**, **510** and **512**, together delimited by box **514**, scores are set to zero when, due to background correction and normalization, it cannot be said accurately whether the direction of the relative score is real. At steps **516**, **518**, and **520**, together delimited by box **522**, scores are set to zero when, although not artifactual, they may not be distinguishable statistically from zero.

Continuing on a gene-by-gene basis, a final manipulation **524** corrects for the disparate dynamic ranges of gene expression manifested by the various genes of the organism. For example, some genes may be capable of only a two-fold change in gene expression no matter how severe the change in condition; other genes may be capable of a 200-fold change in gene expression. To prevent those genes with greater dynamic range from unduly skewing the comparative analyses, each relative expression score is divided by the log of the square

root of the historical maximum expression observed for that gene over all prior experiments. As shown at 524, each relative expression score is divided by the log square root of the largest signal historically output from step 108; that is, each relative expression score is divided by the log square root (one-half the log) of the largest normalized signal observed historically for that gene. As will be understood by those skilled in the art, the value for each gene will depend both upon the expression matrix technology (such as array size) and the data previously collected, and will, on occasion, change as further experiments are done.

Alternatives exist to account in step **524** for the disparate dynamic ranges of the various genes.

In one such alternative, each relative expression score is divided by the log square root of the largest signal historically output from step 108 - that is, by the largest normalized signal — with the difference from the first approach lying in the value chosen to accomplish the normalization ("Σ Signals" in step 108). This approach is further discussed and exemplified in Example 5, below.

In yet another alternative, each relative expression score is divided by the log square root of the largest signal historically input to step 108; that is, each relative expression score is divided by the log square root (one-half the log) of the largest un-normalized signal observed historically for that gene. This may be particularly preferred in circumstances in which normalization proves inappropriate.

Alternatively, one can divide by the magnitude of the largest log signal — either normalized or unnormalized — rather than dividing by its log square root. A rationale for selecting the square root

of the largest signal in the present method is that certain types of errors vary as the square root of the signal. The log of the square root correction was found to yield more informative expression profile comparisons.

A further alternative approach is to make no correction at all, on the assumption that genes whose expression can vary the most are biologically more important, or at least more significant in assessing relatedness of environmental conditions.

Yet another alternative treats the various genes differently, depending upon their empirically-determined significance to the analysis being performed. For example, most of the genes may be treated as above-described, dividing by the log of the square root of the historical maximum expression observed for that gene over all prior experiments. A predetermined subset of particular genes, however, may be differentially treated at this step to increase or decrease their significance in the subsequent analysis.

The aforementioned steps, collectively delimited by box **528**, are followed for each of the genes commonly represented in a first and second gene expression profile. For some expression matrices such as those that measure gene expression in prokaryotes or small eukaryotes such as yeast, all, or substantially all, open reading frames may be so compared. For other platforms using mammalian cells, a large number, and perhaps a comprehensive number, of genes will be assessed. Clearly, only those genes commonly measured as between a first and a second environmental condition can be used to generate a relative gene expression score.

A final, scalar measure, also termed a composite score, which expresses in a scalar value the relatedness of the gene expression profiles of the two conditions, may be calculated **526** by summation. The lower the resulting number, the more closely related the gene expression profiles under the two compared conditions, with complete identity giving a value of zero.

Although no further correction is required, the summation is optionally and preferably corrected **526** for the percentage of genes that usably contributed to the score.

The percentage of genes that prove unusable, that is, that are removed at the steps delimited by box **514** through assignment of their relative Score to zero, **508** and **512**, has an effect on the composite score. Thus, in the optional correction for unusable genes **526**, the simple summation of relative expression Scores is further multiplied by the ratio given by the number of genes divided by usable genes.

The analyses presented in Examples 1 - 4 below were performed on gene expression profiles acquired from matrices with 864 reporters. Although not so indicated in **FIG. 5**, the scores obtained from step **526** may optionally be normalized to express the relative expression score per 1000 genes, to permit comparisons from different sized matrices. To accomplish this normalization, the relative profile score **526** is further multiplied by the ratio of 1000 divided by the total number of genes in the matrix used.

The above-described method allows one quantitatively to rank the relatedness of two gene expression profiles: the lower the resulting composite score, the more related the profiles; the more related the profiles, the more related the global gene expression state of the cells under the two distinct conditions under which the gene expression profiles were obtained.

Thus, one may assess quantitatively the relatedness of two environmental conditions on the global gene expression of a cell. The environmental condition may, for example, be incubation in different media, as further demonstrated in Example 4 below. Alternatively, the two environmental conditions may comprise treatment with two different chemicals, such as pharmaceutical drug candidates, with the relatedness of the gene expression profiles, as reported by the composite score, indicating the relatedness of the action of the drugs. This aspect of the invention is demonstrated in Examples 1 - 3.

The method may also be used quantitatively to relate a preselected environmental condition to a defined genetic mutation of a cell, comprising the steps of: (a) obtaining a first gene expression profile from a cell bearing a mutation and obtaining a second gene expression profile from a wild-type cell under a preselected environmental condition; and then (b) quantifying the relatedness of the first and second gene expression profiles.

In a preferred embodiment of this aspect of the invention, the environmental condition under which expression data are acquired from the wild type cell comprises exposure to a chosen chemical compound. Beginning with a defined mutation, this approach allows

one quantitatively to identify drug candidates that mimic, in their effect, the genetic mutation. Conversely, starting with the gene expression profile of an important pharmaceutical agent, mutations that mimic the effects of the drug may be identified by the quantitative relatedness of their gene expression profile to that obtained in the presence of the drug. The result is the elucidation of the mechanism of drug action through identification of all targets, direct and indirect, affected by the drug. Furthermore, the relatedness of two mutations may be determined by quantitatively relating the gene expression profile obtained from each, absent additional drug.

In applications of the quantitative methods of the present invention to analysis of genetic mutations, the cells are preferably yeast cells, more preferably, *Saccharomyces cerevisiae*. Yeast are particularly preferred for this purpose, and for other applications in which relatedness of genetic mutations is assessed, because (1) the entire genome of *S. cerevisiae* has been sequenced, (2) targeted deletions or insertions may readily be made by homologous recombination, and (3) many fundamental metabolic pathways are highly conserved as between yeast and humans. See, *e.g.*, the discussion in Lashkari *et al.* The methods may be applied more broadly, however, whenever mutations are identified in the cells of other prokaryotic or eukaryotic organisms.

Although the description given above has referred particularly to a method for relating quantitatively a first and second gene expression profile, the present invention also provides a method

for ordering the relatedness of a plurality of gene expression profiles.

To accomplish a ranking which orders the relatedness of a plurality of gene expression profiles, a series of composite scores are obtained, each measuring the relatedness to a common index, or reference, profile. Thereafter, the composite scores are ordered, with lower scores indicating greater relatedness to the index profile. Such ordered rankings are presented in the Tables below.

Thus, the invention provides a method to order the relatedness of environmental conditions to a single preselected environmental condition upon a cell, comprising the steps of: (a) obtaining from the cell or from genotypically identical cells a gene expression profile for each of the plurality of environmental conditions and for the preselected environmental condition; (b) quantifying pairwise the relatedness of each of the plurality of gene expression profiles to the preselected gene expression profile; and (c) ordering these pairwise-measured quantities. In a preferred embodiment, one or more of the environmental conditions comprises exposure of the cells to a chemical compound.

Analogously, the invention also provides a method to order the relatedness of each of a plurality of environmental conditions to a defined genetic mutation of a cell, comprising the steps of: (a) obtaining a set of first gene expression profiles from a wild type cell under each one of the plurality of environmental conditions and a second gene expression profile from a cell having said defined mutation; (b) quantifying pairwise the relatedness of each of said first gene expression profiles to said

second gene expression profile; and then (c) ordering the pairwise-measured quantities.

In like fashion, the invention also provides a method to order the relatedness of each of a plurality of genetic mutations to a defined, or preselected, mutation of a cell, comprising the steps of: (a) obtaining a set of first gene expression profiles from cells each having one of the plurality of genetic mutations and a second gene expression profile from a cell having the preselected mutation; (b) quantifying pairwise the relatedness of each of the first gene expression profiles to the second gene expression profile; and then (c) ordering said pairwise-measured quantities.

Method for Quantifying the Relatedness of Gene Expression Profiles By Linear Regression

The composite score, and thus the ranking of relatedness that is provided by the procedures of FIG. 5, is weighted substantially by outliers, that is, by those genes whose expression changes substantially as between the two measured conditions. This is true notwithstanding the correction for the dynamic range of expression of the various genes, **524**, and results from steps **516**, **518**, and **520**, delimited by box **522** in FIG. 5, in which application of a threshold requirement for data inclusion reduces the contribution by genes with small changes in expression as between the measured conditions. An advantage of such bias is that it focuses the ranking on genes that contribute most substantially to the phenotypic change.

FIG. 6 provides an alternative method for quantitatively relating gene expression profiles, one

that instead weights the ranking of relatedness more toward the commonality of the direction of change in individual gene expression, rather than the magnitude of such change. The method presented in FIG. 6 presents several advantages over that set forth in FIG. 5, particularly the ability accurately to relate gene expression profiles obtained using small concentrations of pharmaceutical agents, and is now preferred for quantitating the relatedness of profiles acquired under mild treatment conditions, such as low concentrations of drug. The method of FIG. 5, however, remains preferable for quantitating the relatedness of gene expression profiles acquired under more severe treatment conditions, such as treatment with high concentrations of drugs. The choice as between using the algorithm set forth in FIG. 5 or that set forth in FIG. 6 is one that may be made empirically after comparison of the results; such choice is within the skill in the art.

Before discussing the details of this alternative method, the conceptual difference between the two methods may best be visualized by considering the scatter plot of FIG. 2. As noted above, FIG. 2 represents as a scatter plot the relative gene expression of distinct genes in yeast cells that have been treated individually with two closely related antineoplastic chemotherapeutic drugs. As discussed above, the treatments are seen to be closely related, each affecting both the direction and the magnitude of individual gene expression equivalently: as a result, most of the points lie approximately on a line through the origin. It will be understood that identical conditions, absent background, absent noise, and absent other variation, would produce theoretically a series

of expression points all of which lie exactly on a line through the origin.

The threshold applied in steps **516**, **518**, and **520** (delimited by box **522**) in FIG. 5 may be conceptualized, in FIG. 2, as two parallel lines of identical slope equidistant from the regression line drawn through the data, somewhat akin to a confidence interval. The lower the threshold applied empirically in step **516**, the more closely the threshold lines may be conceived to lie to the data regression line, and the greater the number of data points that lie outside; the higher the threshold applied empirically in step **516**, the further the threshold lines may be conceived to lie from the data regression line, and the fewer the number of data points that lie outside. Because only those points that lie outside the threshold lines contribute to the expression profile score (compare step **518** to **520**), the method set forth in FIG. 5 is affected substantially by the distance such points lie from the regression line.

The method set forth in FIG. 6, by contrast, focuses on the degree to which the data points fit the theoretically perfect regression line that signifies identity in treatments. Those points that fall directly on the regression line, rather than being least significant in the analysis, contribute substantially to the score. Rather than asking the magnitude of change in gene expression, this method focuses instead on the direction of changes in gene expression. This method proves less sensitive to the concentration of the various drug treatments being compared, as shown below in Example 3.

FIG. 6 schematizes this second approach to quantifying the relatedness of two gene expression profiles.

The gene expression signal for each gene commonly represented in the first (Signal1 **600**) and second (Signal2 **601**) gene expression profile, as processed according to FIG. 1, is input. The Signals have been further corrected for matched controls according to the algorithm set forth in FIG. 1B.

Next, a manipulation **610, 611** — analogous to that performed at step **524** in the earlier algorithm set forth in FIG. 5 — corrects for the disparate dynamic ranges of gene expression manifested by the various genes of the organism.

The same alternatives for adjusting for dynamic range as are set forth above with respect to step 524 apply here as well. Thus, Signal 600, 601 may be divided by the log square root of the maximum (normalized) signal historically output from step 108; may be divided by the log square root of the maximum signal historically input to step 108; may be divided by the log square root of the maximum (unnormalized) signal historically input to step 108; may be divided by the log of the maximal signal — either normalized or unnormalized — rather than by the log square root; may be left unaltered, making no correction for dynamic range at all; or may be adjusted individually using empirically chosen values. A further alternative, exemplified and further discussed in Example 5, below, adjusts the dynamic range of all of the genes in the analysis by dividing by the log square root of the maximum normalized value, but with the value used for normalization chosen from a larger set of genes.

Next, the first (Signal1 **610**) and second (Signal2 **611**) expression signal are associated **620** to provide, for each gene, two-dimensional coordinates. Linear regression **625** on the collection of paired data — representing the expression of all genes commonly represented in the two gene expression profiles — then provides a Score **626** that provides a quantitative measure of the relatedness of the two gene expression profiles, with higher numbers indicating a closer degree of relatedness. The correlation coefficient itself may be used as the score, as may be any multiple thereof. The scores provided in the Examples below were further derived by multiplying the correlation coefficient by 100.

Thus, where the first algorithm (FIG. 5) collapses the first and second expression signal for each gene into a single scalar value **504** (representing a ratio of expression as between first and second gene expression profiles) before summing these values to obtain a composite score, the present algorithm retains the values as separate coordinates until the final step.

It will be understood that any data structure that permits the first and second signal for each commonly represented gene to be associated for purposes of linear regression may be used, such as a single 2-dimensional matrix, a set of vectors, or the like. It will further be understood that any statistical method that reports the closeness of the fit of the data to a best-fit theoretical line through the two-dimensional data may be used according to this invention for calculation of the relative profile score in steps **625** and **626**. Those skilled in the art are both able to

identify such data structures and statistical methods and to encode such calculations in a digital computer; it is the discovery that such closeness of fit permits reliable, reproducible, and ready quantitation of the relatedness of gene expression profiles that is newly described herein.

An additional step, not described in FIG. 6, may optionally be added to the present method.

Signal1 **600** and Signal2 **601** may be subjected to queries identical to those presented at **506** and **510**. That is, the question may be posed whether the earlier background correction and normalization potentially precludes the definitive determination of the direction of change in expression as between the two conditions. If so, that is, if the query presented at either **506** or **510** returns true, the Signals for the gene may optionally be omitted from the linear regression.

The method described in FIG. 6 may be used, like that set forth in FIG. 5, to assess quantitatively the relatedness of two environmental conditions on the global gene expression of a cell; to assess quantitatively the relatedness of a preselected environmental condition to a defined genetic mutation of a cell; and to quantify the relatedness of two different mutations. Further, the algorithm and methods set forth in FIG. 6 may be used, like that set forth in FIG. 5, to order the relatedness of a plurality of gene expression profiles, whether acquired under disparate environmental conditions, acquired from cells bearing various mutations, or acquired from a combination thereof.

As presented above, each gene commonly represented in a first and second gene expression profile is treated identically to other genes represented in the gene expression profiles, whether the algorithm given in FIG. 5 or that given in FIG. 6 is applied. However, it is possible — and will often be recommended — differentially to weight changes in the expression of one or more preselected genes, so as to increase or decrease their significance in the analysis. Such weighting may be done, for example, by adjusting the Signal at step **524** or at step **610**, **611**.

Data Storage

For each embodiment of this invention, whether using the method described in FIG. 5 or that described in FIG. 6, data may be stored for any individual gene expression profile at any or all of the intermediate points in the processes described in FIGS. 1, 5, or **6**. Data acquired from any single expression matrix may, for example, be stored as raw digitized data as obtained at step **101**, as background-adjusted, normalized signals as obtained at step **108**, as the log of background-adjusted, normalized signals as obtained at step **110**, or as signals fully corrected for matched controls, as obtained in step **112**.

It will be appreciated that new comparisons for relatedness -- that is, new calculations of composite scores according to the algorithm of FIG. 5 or the calculation of relative profile scores according to the algorithm of FIG. 6 -- may be performed using data that were earlier acquired and stored. Thus, as additional experiments are run and additional profile

data acquired from the various gene expression matrix platforms described herein, more data become available for the described analyses. In particular, as more drugs are tested for their effects on global gene expression, an increasingly comprehensive database will be established from which comparisons may be made.

The storage of gene expression profiles, each representing a distinctive cellular state to which reference may repeatedly be made for purposes of comparison, is analogous to the compilation of spectra identifying discrete states of inanimate matter -- NMR spectra, IR spectra, mass spectra, and the like -- comparison to which standards allow the identification of unknown chemical structures. Comparisons of gene expression profiles may be used in like fashion. Conversely, the quantitative assessment of relatedness provided by the methods and apparatuses described herein may be applied to such other spectra, with modifications as would be well understood by those skilled in the art.

Drug Discovery And Other Uses For Quantitative Analyses of Gene Expression Profiles

The quantitative methods, systems, and apparatuses provided herein permit new drug discovery approaches. By quantifying the relatedness of gene expression profiles, compounds may be tested for similarity to drugs of known mechanism, to drugs of known efficacy, or for similarity to defined mutations, conditions, disorders or disease states.

Treatment of a target cell with a drug, no matter what the primary biologic process perturbed by that chemical, results ultimately in a change in the

pattern of gene expression in the target cell. Drugs which act similarly produce similar patterns of change. The greater the similarity in action, the greater the similarity in the change in gene expression profiles. As a consequence, the ability to quantitate the relatedness of gene expression profiles may permit the identification of drugs with similar global effects on cellular gene expression; drugs, by inference, which have similar mechanisms of action.

When the mechanism of action of a first drug is known, identifying other chemical compounds that effect similar changes in the gene expression profile of a target cell may identify additional compounds sharing similar mechanisms of biological action. When the mechanism of a first drug is not known but the drug is known to be effective in treating a given disorder, identifying drugs that effect similar changes in the gene expression profile of a target cell may identify drugs that are similarly effective in treating that pathologic state, albeit drugs of similarly unknown mechanism.

Thus, the ability to quantitate the relatedness of gene expression profiles may obviate the present need to identify an isolated pharmaceutical target, to develop a dedicated assay, and then to screen compounds for their activity in the dedicated assay.

The ability to quantitate the relatedness of gene expression profiles may, moreover, facilitate efforts during latter stages of drug development to narrow and focus the specificity of action of promising drug candidates. For example, pharmacologically-effective derivatives of a lead compound may be identified, as above, based on quantitative relatedness

of their gene expression profiles to that of a lead candidate.

The experimental Examples that follow demonstrate some of these applications of the quantitative methods of the present invention.

In Example 1, the relatedness of drugs to actinomycin D was assessed by quantitative comparison of a gene expression profile obtained in the presence of actinomycin D to a plurality of gene expression profiles obtained upon exposure to other pharmaceutical agents. Using either of the above-described algorithms, varying concentrations of daunarubicin, 5-FUDR, doxorubicin, 5-FU, hydroxyurea and mycophenolic acid were identified as causing quantitatively similar effects on the global gene expression of the cell, here an *S. cerevisiae* cell. All of these agents, like actinomycin D, are known to affect nucleic acid synthesis.

Thus, were the mechanism of action of actinomycin D alone known, the data would clearly implicate daunarubicin, doxorubicin, the nucleotide analogues 5-FUDR and 5-FU, and mycophenolic acid as drugs with mechanisms of action similar to the known mechanism of actinomycin D. Knowing that actinomycin D interferes with nucleic acid synthesis, the data indicate that daunarubicin, doxorubicin, the nucleotide analogues 5-FUDR and 5-FU, and mycophenolic acid also affect nucleic acid synthesis, and may, therefore, be useful as chemotherapeutic agents in the treatment of cancer, or may have utility in interrupting the life cycle of pathogens, particularly viral pathogens.

Conversely, were the mechanism of all of these agents but the reference drug known, these data would indicate that actinomycin D interferes with

nucleic acid synthesis, providing valuable insight into its mechanism.

It should be noted that these insights did not require a dedicated nucleic acid synthesis inhibition assay, nor prior identification of a molecular target for the drugs. And as a result, drugs with similar global effects, but disparate molecular targets, have been identified.

Examples 2 and 3 similarly assess the relatedness, as measured by changes in global gene expression, of a plurality of drugs to one of two concentrations of daunarubicin, again demonstrating that the relatedness of action can be determined without foreknowledge of the structure or mechanism of the preselected reference drug. Example 4 demonstrates that the methods set forth herein may be used more broadly, quantitatively to relate the effects on a cell of global environmental conditions.

Methods of Selecting Informative Gene Subsets For Gene Expression Profiling

The gene expression profiles that are quantitatively compared in the analyses presented in Examples 1 - 4 each contains data on the contemporaneous level of expression of over 800 different *S. cerevisiae* genes. These 800 genes represent a subset of the organism's expressible genes, estimated to be just slightly over 6000 in number. The results thus demonstrate that only a portion of a cell's global gene expression need be assayed for successful application of the methods described herein. Although the quantitative analysis will be increasingly robust and informative as the percentage of assessed

genes increases, it is clear that the expression of fewer than all genes may be used in these analyses.

Often, technical considerations in the acquisition of gene expression data will dictate that fewer than all expressible genes be assayed. For example, samples of drug candidates may be in limiting supply, particularly when produced in small quantity by combinatorial chemistries; there may simply be too little of the agent to permit the testing of its effects on all possible genes of a given cell type. It may also, or in the alternative, be too expensive to assay each candidate agent across each expressible gene of the cell.

These issues are compounded when the genome to be assayed becomes more complex. Thus, to assess the effect of a drug or other environmental agent on each of the expressible genes of a nematode, such as *C. elegans*, would require the measurement of the expression of nearly 20,000 genes; to assess the effect of a drug or other environmental agent on each of the expressible genes of a human cell would require the measurement of about 100,000 genes.

Furthermore, not all genes prove equally informative. Some may have an insufficient dynamic range in expression to provide significant information, no matter what the environmental condition. Other genes may vary in expression coordinately, or cooperatively, providing redundancy in the information collected.

One approach to selecting informative subsets of genes for expression analysis is to choose the genes individually by known or suspected function. Thus, Farr *et al.*, U.S. Patent No. 5,811,231 and European

patent no. EP 0680517 B1 disclose, *inter alia,* the selection of "stress genes" particularly to identify and characterize compounds that are toxic to the cell.

Such an approach, however, requires antecedent knowledge of the gene's function. Furthermore, the bias imposed by such directed selection would reduce the possibility of identifying previously unsuspected relationships; in a method useful for the identification of such unsuspected relationships, such as the methods presented herein, such directed preselection would be particularly disfavored.

Another approach is to choose the subset entirely at random, in the hope that the subset so selected proves representative of the whole. The problem, clearly, is that the subset so chosen may in fact prove uninformative for describing the cellular state under one or more environmental conditions.

Yet another approach would be to select genes identified not by common function, but by a common responsiveness to a preselected environmental condition. Whitney *et al.*, Nat. Biotechnol., 16:1329-33 (1998). Falling somewhere between the purely directed and purely random approach, this latter procedure is, to some extent, subject to the disadvantages of both.

FIGS. 7 and 8 demonstrate qualitatively the results of a novel alternative for selection of informative gene subsets for gene expression analysis, to be described more fully below. This novel approach predicates the selection of genes for expression analysis upon the diversity — rather than size, direction, or commonality — of their expression.

FIG. 7 is a scatter plot of gene expression signals, processed according to FIG. 1, derived from genome reporter matrices comprising 1532 separate *S. cerevisiae* gene expression reporters, each matrix treated individually with one of two agents known to be closely related in structure and function: 10 µg/ml Lovastatin (X axis) and 20 µg/ml Mevastatin (Y axis). As earlier discussed with respect to FIG. 2, it is readily apparent from casual inspection of the Figure that the two drugs affect the expression of most yeast genes similarly, if not identically: each gene whose expression is increased by Lovastatin is equivalently augmented by Mevastatin; each gene whose expression is decreased by treatment with Lovastatin is equivalently repressed by treatment with Mevastatin; and those genes whose expression is unaffected by treatment with Lovastatin are similarly unaffected by Mevastatin. The result is that most of the data points lie on a line through the origin.

FIG. 8 plots the gene expression signals from a 96 gene subset selected from the 1532 gene expression signals presented in FIG. 7. Although only 1 in 16 of the genes presented in FIG. 7 is selected for display in FIG. 8, the strong correlation in the two drug treatments may still be seen. The 96 genes in the selected subset are listed in Table 9, presented in Example 5 below. Although selected without regard to known function, the genes retained in the subset are seen to have diverse functions (the gene functions listed in the Table are drawn from the Stanford University *Saccharomyces* genome data base http://genome-www.stanford.edu/Saccharomyces).

The subset of genes displayed in FIG. 8 was selected from those displayed in FIG. 7 in a process

comprising two basic algorithmic steps: in a first step, each of the genes displayed in FIG. 7 was sorted according to its maximal historical dynamic range of expression; in the second step, an iterative process eliminated from the sorted list all but the first in each group of genes whose expression is strongly correlated. The result is retention in the chosen subset of the diversity of gene response seen in the original set, with each group of correlated genes being represented in the retained subset by that one gene with greatest dynamic response.

Although the principle is exemplified in FIG. 8 by selection of a subset of genes from amongst the larger number of genes for which expression data have previously been acquired, the approach finds greatest utility in directing the prospective acquisition of a smaller, yet informative, number of gene expression signals from the gene expression matrix itself.

Examples 1 - 4 demonstrate that the measurement of the expression of 864 of the 6000 genes potentially expressible by S. cerevisiae — that is, just about 14.4% of the total number of genes potentially expressible by the cell — permits the quantitative definition of cellular phenotype, and thus the quantitative determination of the relatedness of cellular states. Example 5 demonstrates that it is possible to select an even smaller subset of potentially expressible genes — just 96 of 6000, or about 1.6% of potentially expressible genes — the expression of which is sufficiently informative as to permit the quantitative definition of cellular phenotype, and thus the quantitative determination of the relatedness of cellular states.

Thus, an important aspect of the present invention is to provide methods of cellular phenotyping, comprising selecting no more than 20% of a cell's expressible genes for expression analysis, wherein the concurrent expression of the selected genes sufficiently defines the cell's phenotype as to permit the cell's phenotype quantitatively to be related to the phenotype of another cell. In these methods, preferably no more than about 20% of the cell's potentially expressible genes are selected, more preferably no more than about 15% of the cell's potentially expressible genes, even more preferably no more than about 10% of the cell's potentially expressible genes, optimally no more than about 5% of the cell's potentially expressible genes, and in the most preferred embodiments, about 1% - 5%, and even 1 - 2% of the cell's potentially expressible genes. Algorithms for effecting such selection, and computers, systems, networks, and other devices for effecting the methods are also presented.

The two basic steps in the algorithm for selecting an informative subset of expressible genes for expression analysis may be better understood by particular reference to **FIGS. 9** and **10.**

The first of two major steps in the algorithm orders genes according to the dynamic range of their expression. Preferably, historical data are used: for each gene, the maximum and minimum value of Signal 108 in the database of electronically stored gene expression profiles is determined by an appropriately formulated query (or series of queries) **900.**

As noted above, gene expression data may be stored at any or all of the intermediate points in the processes described in FIGS. 1, 5, or 6. For purposes

of the algorithmic steps set forth in FIG. 9, the Signal as output from step **108** is used. If Signal values as output from step **108** are not present in the database, the values may in certain instances be reconstructed from the values so stored-- for example, if the Signal values output from step **110** are stored, the Signal as it would have been output from step **108** may be calculated by reversing step 110, that is, by exponentiation.

The Range of expression is calculated **902** as the ratio of maximum to minimal signal (assign Range = $Signal_{max}/Signal_{min}$). Although other measures of dynamic range may be used — such as "$Signal_{max}$ - $Signal_{min}$" — the ratio is presently preferred.

Next, a threshold is applied **904** by comparing the Range obtained in step **902** to a value that is empirically established. If Range exceeds the threshold, the gene is retained for subsequent use; if Range fails to exceed the threshold, the gene is discarded from further analysis. As shown in step **906**, that discard may readily be achieved by setting Range to a null value. For the selection shown in FIG. 8 and exemplified in Example 5, a threshold of 10 was set. That is, only those genes demonstrating at least a 10-fold change in gene expression level across the set of historical gene expression profiles stored in the database were retained in the selected subset.

The selection of a range threshold at this step in the algorithm will be determined by empiric needs, and is well within the skill in the art. Typically, a threshold of 10-fold will provide informative subsets of appropriately reduced size.

It is, however, possible to set the threshold as low as 1; that is, to eliminate the cutoff entirely. The result, all other factors held constant, will be the selection of a much larger subset of genes. Furthermore, it will be understood that the threshold that is set at this step need not be limited to whole numbers.

Thus, the threshold may be set as low as 1 or may, preferably, be greater than 1. Usually, the threshold will be set at 2 or greater, more preferably at 3 or greater, even more preferably at 4, 5, 6, 7, 8, or 9 or greater, in that order, most preferably to at least 10.

The threshold may also be greater than 10, ranging as high as 100, preferably no more than 50, more preferably no more than 25, most preferably 10 - 20.

The genes whose range of expression exceeds the empirical threshold are then assorted according to expression Range.

FIG. 10 schematizes the second, iterative process of the second basic algorithmic step.

Proceeding from left to right, FIG. 10 outlines two full iterations of the second step of the algorithm. At the left is shown the list of genes, as output from step 908, ordered from greatest to least dynamic range. Genes that were discarded at step 906 due to inadequate dynamic range are not shown.

In the first iteration of the process, the first gene in the list ("gene 1") serves as the index, or reference, gene. Taking each successive gene in the list in turn, the degree to which that gene's expression is correlated with the expression of the index gene across the set of stored gene expression

profiles, is calculated. If the correlation (r^2) exceeds an empirically set value, the gene is discarded from the set.

The effect of this step is to remove all genes whose expression is strongly correlated with that of the index gene, "gene 1"; the high degree of correlation implies that information contributed by the expression of these discarded genes is in large measure redundant of the information inherent in the expression values of the index gene. As shown at the bottom of FIG. 10, the index gene ("gene 1") is retained in the informative gene subset; as exemplified in the middle of FIG. 10, genes highly correlated therewith ("gene 3" and "gene 4") are discarded. Because the list is ordered from greatest to least expressive range, the single gene retained from the correlated group is that with the greatest dynamic range of expression.

In the second iteration of the process, the first of the genes retained after gene 1 (exemplified by "gene 2" in FIG. 10) becomes the index, or reference gene. It too will be retained, as shown at the bottom of the figure.

Taking, in turn, each successive gene that has been retained in the list, the degree to which that gene's expression is correlated with the expression of the index gene (now "gene 2") across the set of stored gene expression profiles, is calculated. If the correlation (r^2) exceeds an empirically set value, the gene is discarded from the set. The next retained (uncorrelated) gene, here exemplified by "gene 6", then becomes the index gene for the next iteration.

The process is repeated until the list is exhausted.

In performing the iterative step of removing genes whose expression is correlated with that of the index gene, the correlation is preferably performed on the gene expression Signal as output from step 140 (*i.e.*, as output from box 141). The number of genes retained in the final subset will be determined by the total number of genes contributing data to the database of gene expression profiles, by the range threshold applied at step 904, and by the correlation threshold applied during the iterative process schematized in FIG. 10. The two threshold values may be adjusted empirically to yield an informative subset containing any chosen number of genes.

Thus, in the analysis presented below in Example 5, the range threshold and correlation threshold were adjusted empirically to provide a subset with 96 genes — equal to the number of wells of standard microtiter plate — by setting the range threshold to 10 and the correlation threshold to 0.675.

Once the subset of desired size is identified according to the algorithm set forth in FIGS. 9 and 10, quantitative analyses may be performed, using just that subset of genes, according to the algorithms set forth in FIGS. 5 and 6. The analyses may be performed, as in Example 5, by selecting from more comprehensive gene expression profiles, or may, more usefully, be performed by acquiring prospectively gene expression profiles using just the identified subset of genes in the reporter matrix.

Example 5 demonstrates the selection of a subset of 96 genes from the 1532 genes available in our database of stored gene expression profiles. A comparison of the data in Tables 8 and 10 — Table 8 ordering the relatedness using 1532 genes, and Table 10

ordering the relatedness of the same profiles based upon just 96 genes selected using the described approach — demonstrates that the 96 gene subset retains sufficient diversity to permit the quantitative ordering of the relatedness of gene expression profiles: the data in both Tables identify HMG-CoA reductase inhibitors as most closely related to Lovastatin, with drugs that affect other steps in the sterol biosynthetic pathway as next most closely related in effect.

Although the quantitative analysis of gene expression in Example 5 was performed on the 96 gene subset using the algorithm of FIG. 6 (*i.e.*, FIGS. 1A, 1B, and 6), the algorithm given in FIG. 5 (*i.e.*, FIGS. 1A, 1B, and 5) may also be used. Furthermore, FIG. 8 — which plots the expression data for the 96 genes from the index profile (appearing at rank 0) versus data from the profile appearing at rank 2 (20 µg/ml Mevastatin in 1% Ethanol) — demonstrates that the subset so selected may also be used for the qualitative analysis of gene expression profiles.

The following examples are offered by way of illustration and not by way of limitation.

EXAMPLE 1

Relatedness of Drugs to 80 µg/ml Actinomycin D

Replicate genome reporter matrices were prepared according to Ashby *et al.*, which is incorporated herein by reference. Briefly, for each such matrix recombinant constructs, each driving a

fluorescent reporter from a distinct yeast promoter, were transformed individually into discrete cultures of *Saccharomyces cerevisiae* of identical strain background. Selection was applied to transformed cultures both to maintain the reporter and to prevent contamination by untransformed cells. Each such culture of transformed yeast was segregated and maintained in a separate spatially-addressable well of the matrix.

The matrices as used contained 864 separate constructs, permitting the contemporaneous measurement of the expression levels of over eight hundred genes. Each matrix was subjected to a defined environmental condition, as specified in the entries of Tables 1 and 2. A gene expression profile was obtained from each matrix, as set forth in Ashby *et al.*, digitized, and stored electronically.

Thereafter, the relatedness of each gene expression profile to that produced in the presence of 80 µg/ml actinomycin D was quantified pairwise, substantially according to the method set forth in FIGS. **1A**, **1B** and **5** (Table 1), or as set forth in FIGS. **1A**, **1B**, and **6** (Table 2). The measures of pairwise relatedness were then ordered, with the following results:

TABLE 1

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
0	80 Actinomycin D in 1% methanol (index, or reference, condition)	0

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
1	60 Actinomycin D in 1% methanol	2.9
2	40 Actinomycin D in 1% methanol	10.0
3	50 Actinomycin D in 1% methanol	11.7
4	25 Daunarubicin	14.2
5	50 Daunarubicin	15.6
6	40 5-FUDR	15.8
7	25 Doxorubicin	16.0
8	12.5 Doxorubicin	16.0
9	25 Doxorubicin	17.7
10	30 FUDR	18.0
11	12.5 Doxorubucin	21.2
12	0.30 FUDR	21.9
13	5000 Hydroxyurea	22.3
14	20 5-FUDR	22.4
15	0.1 5-FU	22.5
16	12.5 Daunarubicin	22.9
17	0.25 5-FU	23.0
18	6.25 Doxorubicin	23.0
19	30 Actinomycin D in 1% Methanol	23.5
20	9 Mycophenolic acid in 1.5% Ethanol	25.1
21	40 Actinomycin D in 1% Methanol	26.8
22	0.250 5-FU	27.7
23	15 Mycophenolic Acid in 1.5% Ethanol	28.1
24	2 Flucytosine (15 hr)	28.1

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
25	0.15 5-FU	28.4
26	5 Alpha factor	32.1
27	10 Alpha factor	32.2
28	50 Mevastatin in 2% DMSO	38.2
29	75 Mevastatin in 2% DMSO	38.4
30	20 Alpha Factor	40.6
31	No Drug in 1% Methanol	41.1
32	0.04 Miconazole in 1% DMSO	46.3
33	100 Mevastatin in 2% DMSO	55.5
34	250 Griseofulvin in 1% Methanol	56.5
35	15 Alpha Factor	66.7
36	4000 Verapamil	92.0
37	3500 Verapamil	113.1
38	4500 Verapamil	141.1
39	0.08 Miconazole in 1% DMSO	158.8
40	0.156 Sulconazole in 1% DMSO	169.7

TABLE 2

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
0	80 Actinomycin D in 1% Methanol (index, or reference, condition)	100
1	60 Actinomycin D in 1% Methanol	86
2	50 Actinomycin D in 1% Methanol	74
3	40 Actinomycin D in 1% Methanol	72
4	25 Doxorubicin	68
5	40 5-FUDR	67
6	25 Daunarubicin	65
7	12.5 Daunarubicin	65
8	50 Daunarubicin	65
9	0.3 5-FU	64
10	30 5-FUDR	63
11	0.25 5-FU (expt. 641)	62
12	0.25 5-FU (expt. 351)	62
13	0.35 5-FU	60
14	25 Doxorubicin	59
15	50 Doxorubicin	59
16	0.2 5-FU	59
17	6.25 Doxorubicin	58
18	0.1 5-FU	58
19	12.5 Doxorubicin	53
20	12 Mycophenolic Acid in 1.5% Ethanol	53

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
21	5000 Hydroxyurea	52
22	9 Mycophenolic Acid in 1.5% Ethanol	51
23	12.5 Daunarubicin	49
24	10000 Hydroxyurea	49
25	15 Mycophenolic Acid in 1.5% Ethanol	49
26	2 Flucytosine	48
27	4 Flucytosine (expt. 167)	48
28	4 Flucytosine (expt. 97)	48
29	5000 Hydroxyurea	46
30	2 Flucytosine (15 hrs)	45
31	No Drug in 10% Methanol	42
32	7.5 Alpha Factor	36
33	10 Alpha Factor	36
34	4500 Verapamil	36
35	3500 Verapamil	35
36	20 Alpha Factor	35
37	3000 Verapamil	34
38	4000 Verapamil	33
39	4 Alpha Factor	31
40	1250 Hydroxyurea	30
41	5 Mevastatin in 1% DMSO	28
42	2500 Verapamil	28
43	2 Mycophenolic Acid in Ethanol	28

Tables 1 and 2 demonstrate that each of the methods described herein is able to quantitate the relatedness of gene expression profiles, and by so doing, to identify the relatedness of drug treatments.

Thus, as set forth in Table 1, the algorithm of FIGS. **1A**, **1B**, and **5** identifies treatment with 60 µg/ml actinomycin D as the most closely related of the treatments to the reference, or index, condition, which is exposure to 80 µg/ml actinomycin D. Treatment with 40 µg/ml actinomycin D and 50 µg/ml actinomycin D follow thereafter.

Varying concentrations of daunarubicin, 5-FUDR, doxorubicin, 5-FU, hydroxyurea and mycophenolic acid follow. All of these agents, like actinomycin D, are known to affect nucleic acid synthesis. Much less closely related are treatments with agents of disparate activity: treatment with yeast alpha factor at rank 26 and 27, followed thereafter by Mevastatin, the latter an inhibitor of HMG-CoA reductase. At rank 31 may be found the profile generated by treating with no drug at all, the environmentally-matched control, and below that, treatment with the antifungal agents miconazole and griseofulvin, and treatment with the calcium channel blocker verapamil.

Thus, were the mechanism of action of actinomycin D alone known, the data would clearly implicate daunarubicin, doxorubicin, the nucleotide analogues 5-FUDR and 5-FU, and mycophenolic acid as drugs with mechanisms of action similar to the known mechanism of actinomycin D. Knowing that actinomycin D interferes with nucleic acid synthesis, the data indicate that daunarubicin, doxorubicin, the nucleotide analogues 5-FUDR and 5-FU, and mycophenolic acid also affect nucleic acid synthesis, and may, therefore, be

useful as chemotherapeutic agents in the treatment of cancer, or may have utility in interrupting the life cycle of pathogens, particularly viral pathogens.

Conversely, were the mechanism of all of these agents but the reference drug known, these data would indicate that actinomycin D interferes with nucleic acid synthesis, providing valuable insight into its mechanism.

It should be noted that these insights did not require a dedicated nucleic acid synthesis inhibition assay, nor prior identification of a molecular target for the drugs. And as a result, drugs with similar global effects, but disparate molecular targets, have been identified.

Table 2 presents a quantitative ranking of relatedness of gene expression profiles generated using the method and algorithm of **FIGS. 1A, 1B**, and **6**, as applied to the same set of electronically-stored gene expression profile data.

As can be seen, agents that affect nucleic acid synthesis are again ranked as most closely related to treatment with 80 µg/ml actinomycin D. Of note is the ordered ranking of the decreasing concentrations of actinomycin D.

EXAMPLE 2

Relatedness of Drugs to 50 µg/ml Daunarubicin

Gene expression profiles were obtained and stored as set forth in Example 1 and Ashby *et al.*

Thereafter, the relatedness of each gene expression profile to that produced in the presence of

50 µg/ml daunarubicin was quantified pairwise, substantially according to the method set forth in **FIGS. 1A, 1B** and **5** (Table 3), or as set forth in **FIGS. 1A, 1B,** and **6** (Table 4). The measures of pairwise relatedness were then ordered, with the following results:

TABLE 3

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
0	50 Daunarubicin (index, or reference, condition)	0.0
1	25 Doxorubicin (expt. 336)	2.3
2	50 Doxorubicin	9.7
3	25 Daunarubicin	12.4
4	80 Actinomycin D in 1% Methanol	15.6
5	12.5 Doxorubicin (expt. 335)	17.6
6	60 Actinomycin D in 1% Methanol	19.5
7	0.2 5-FU	24.3
8	0.35 5-FU	24.3
9	40 5-FUDR	25.7
10	6.25 Doxorubicin	26.4
11	0.25 5-FU	26.4
12	12.5 Daunarubicin	26.5
13	0.15 5-FU	26.6
14	40 Actinomycin D in 1% Methanol (expt. 491)	28.9
15	10 Alpha Factor	30.8
16	5 Alpha Factor	30.8

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
17	5000 Hydroxyurea	32.6
18	40 Actinomycin D in 1% Methanol (expt. 456)	33.7
19	2 Flucytosine	35.9
20	20 Alpha Factor	39.9
21	10000 Hydroxyurea	40.7
22	No drug	43.7
23	75 Mevastatin in 2% DMSO (expt. 1202)	43.9
24	1000 Verapamil	44.0
25	20 Alpha Factor	44.1
26	50 Mevastatin in 1% DMSO	44.5
27	75 Mevastatin in 2% DMSO (expt. 1098)	47.6

TABLE 4

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
0	50 Daunarubicin (index, or reference, condition)	100
1	25 Doxorubicin (expt. 336)	91
2	50 Doxorubicin (expt. 337)	90
3	25 Daunarubicin	77
4	12.5 Doxorubicin (expt. 335)	75
5	6.25 Doxorubicin	62
6	0.35 5-FU	59
7	0.2 5-FU	58
8	4500 Verapamil	57
9	60 Actinomycin D in 1% Methanol	57
10	12.5 Daunarubicin	57
11	0.3 5-FU	57
12	0.25 5-FU (expt. 351)	56
13	0.25 5-FU (expt. 641)	56
14	0.15 5-FU	55
15	50 5-FUDR	53
16	12 Mycophenolic Acid in 1.5% Ethanol	52
17	10000 Hydroxyurea (expt. 205)	51
18	4000 Verapamil	50
19	3500 Verapamil	50
20	10000 Hydroxyurea (231)	49
21	15 Mycophenolic Acid in 1.5% Ethanol	49

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
22	No Drug in 10% Methanol	44
23	150 Mitomycin C	43
24	30 5-FUDR	43
25	7.5 Alpha Factor	40
26	3000 Verapamil	40
27	5 Alpha Factor	34
28	15 Alpha Factor	32
29	2500 Hydroxyurea	30
30	2000 Verapamil	24
31	750 Griseofulvin in 7.5% Methanol	15

The data set forth in Table 3 – generated using the method set forth in **FIG. 5** – identifies as agents that are closely related in action to daunarubicin the following: doxorubicin, actinomycin D, 5-FU, and 5-FUDR, consistent with the known activities of these agents. The data set forth in Table 4 – generated using the method set forth in **FIG. 6** – in contrast, are less clear-cut, with verapamil, a calcium channel blocker, appearing as closely related.

Thus, it can be seen that at more severe treatments, here represented by higher concentrations of drug, the method given in **FIG. 5** may be preferable to that given in **FIG. 6**. Example 3, below, demonstrates that the method given in **FIG. 6** will be preferred at lower concentrations of drug.

It should also be noted from the data in this example that replicate gene expression profiles, that is, gene expression profiles obtained in separate

experiments using identical conditions, give data that rank closely with one another, demonstrating the reproducibility of the analysis.

EXAMPLE 3

Relatedness of Drugs to 12.5 µg/ml Daunarubicin

Gene expression profiles were obtained and stored as set forth in Example 1 and Ashby *et al.*

Thereafter, the relatedness of each gene expression profile to that produced in the presence of 12.5 µg/ml daunarubicin was quantified pairwise, substantially according to the method set forth in **FIGS. 1A, 1B** and **5** (Table 5), or as set forth in **FIGS. 1A, 1B,** and **6** (Table 6). The measures of pairwise relatedness were then ordered, with the following results:

TABLE 5

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
0	12.5 Daunarubicin (index, or reference, condition)	0.0
1	5% Saline	1.0
2	1000 Diltiazem	1.3
3	0.25 5-FU	1.9
4	0.2 5-FU	1.9
5	Anaerobic Growth	1.9
6	1000 Verapamil	2.0
7	2 Mycophenolic Acid in Ethanol	2.0

Rank	Treatment (drug concentrations in µg/ml)	Composite Score
8	1187.5 Acetylsalicylic Acid in 1.25% Ethanol	2.1
9	1000 Acetylsalicylic Acid in 1.25% Ethanol	2.1
10	1250 Acetylsalicylic Acid in 1.25% Ethanol	2.2
11	5 Mevastatin in 1% DMSO	2.5
12	10 Amoxicillin in 2% Ethanol	2.6
13	0.04 Tunicamycin in 0.1% DMSO Tris	2.6
14	None	2.9
15	750 Acetylsalicylic Acid in 3% Ethanol	3.0
16	500 Diltiazem	3.1
17	12.5 Doxorubicin	3.6
18	750 Griseofulvin in 7.5% Methanol	3.9
19	7.5 Alpha Factor	4.1
20	5 Alpha Factor	4.2
21	10 Alpha Factor	4.4
22	25 Doxorubicin	13.7
23	20 Alpha Factor	13.8
24	50 Daunarubicin	26.5
25	50 Doxorubicin	62.3

TABLE 6

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
0	12.5 Daunarubicin (index, or reference, condition)	100
1	25 Daunarubicin	76
2	12.5 Doxorubicin	75
3	25 Doxorubicin (expt. 336)	67
4	6.25 Doxorubicin	63
5	12.5 Doxorubicin	58
6	50 Daunarubicin	57
7	60 Actinomycin D in 1% Methanol	52
8	80 Actinomycin D in 1% Methanol	49
9	50 Actinomycin D in 1% Methanol	48
10	40 Actinomycin D in 1% Methanol	48
11	50 Doxorubicin	44
12	9 Mycophenolic Acid in 1.5% Ethanol	43
13	30 5-FUDR	41
14	5 Mycophenolic Acid in 0.9% Ethanol	36
15	1125 Acetylsalicylic Acid in 2% Ethanol	34
16	30 Actinomycin D in 1% Methanol	33
17	No Drug in 10% Methanol	27
18	750 Acetylsalicylic Acid in 3% Ethanol	25

The results presented in Tables 5 and 6 demonstrate the substantial advantage of the second method for quantifying relatedness of gene expression profiles at low drug concentrations.

As shown in Table 5, the first method, that set forth in **FIG. 5**, is unable accurately to quantitate the relatedness of gene expression profiles to that produced in the presence of only 12.5 µg/ml daunarubicin, ranking 5% Saline and 1000 µg/ml diltiazem (a calcium channel blocker) ahead of 5-FU, which itself just precedes anaerobic growth and verapamil in the ranking.

In striking contrast, the same gene expression profile data analyzed according to the method set forth in **FIG. 6** (Table 6) now ranks, as most closely related to treatment with 12.5 µg/ml daunarubicin, treatments with varying concentrations of doxorubicin, which is known to be closely related to daunarubicin in structure and function.

EXAMPLE 4

Relatedness of Global Environmental Conditions

Replicate genome reporter matrices were prepared as in Example 1 and Ashby *et al.*, with 864 distinct elements reporting the contemporaneous expression of 864 different yeast open reading frames. Gene expression profiles were acquired, under the conditions shown below, for each of the matrices, digitized, and stored. Thereafter, the relatedness of each gene expression profile to that produced by

incubation of cells in yeast minimal media was quantified pairwise, substantially according to the method set forth in **FIGS. 1A, 1B** and **5.** The measures of pairwise relatedness were then ordered, with the following results, as set forth in Table 7:

TABLE 7

Treatment	Composite Score
No drug, yeast minimal medium (None/NM)	0.0
No drug, yeast minimal medium plus casamino acids (None/NM + CAA)	37.6
7.5 yeast alpha factor, yeast minimal medium plus casamino acids (7.5 alpha/NM + CAA)	41.7
5 yeast alpha factor, yeast minimal medium plus casamino acids (5 alpha/NM + CAA)	41.8
No drug, yeast minimal medium plus casamino acids (None/NM + CAA)	45.2
No drug, yeast minimal medium plus casamino acids (None/NM + CAA)	45.9
10 yeast alpha factor, yeast minimal medium plus casamino acids (10 alpha/NM + CAA)	46.4
12.5 yeast alpha factor, yeast minimal medium plus casamino acids (12.5 alpha/NM + CAA)	59.4
No drug, yeast minimal medium plus casamino acids, diploid (a/α) strain (None/NM + CAA/diploid)	63.5
15 yeast alpha factor, yeast minimal medium plus casamino acids (15 alpha/NM + CAA)	71.1

Treatment	Composite Score
No drug, YPD medium (None/YPD)	81.6

As shown in Table 7, the quantitative methods provided herein permit one to order the relatedness of global environmental conditions, here represented by changes in nutrient media, just as can be done with discrete treatments with individual drugs.

In addition, these data confirm that changes in media may substantially affect global gene expression, confirming the importance of including a correction for an environmentally-matched control, as set forth in **FIG. 1B.**

EXAMPLE 5

Selection of an Informative Subset of Genes for Quantitative Analysis of Gene Expression Profiles

Replicate genome reporter matrices were prepared according to Ashby *et al.*, which is incorporated herein by reference. The matrices as used for the analyses presented in this Example contained 1532 separate constructs, permitting the contemporaneous measurement of the expression levels of over fifteen hundred genes, about one quarter of the genes expressible by *S. cerevisiae*. Each matrix was subjected to a defined environmental condition, as specified in the individual entries in each of Tables 8 and 10. A gene expression profile was obtained from each matrix, as set forth in Ashby *et al.*, digitized, and stored electronically.

Thereafter, the relatedness of each gene expression profile to that produced in the presence of 10 µg/ml Lovastatin was quantified pairwise, substantially according to the method set forth in FIGS. 1A, 1B and 6, with two minor differences.

First, normalization step 108 was omitted from the analysis of the 96 gene subset because the assumption of constant mean expression may not prove valid as applied to such a small percentage of the cell's genes.

Second, the correction for disparate dynamic range of the reporters was accomplished in steps 610 and 611 by dividing each gene by the log square root of the maximum normalized signal; however, the value used to effect normalization was in each case that value appropriate to the 1532 gene subset.

The measures of pairwise relatedness were then ordered, with the following results.

TABLE 8

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
0	10 Lovastatin in 1% Ethanol (index, or reference, condition; experiment 1538)	100
1	5 Lovastatin in 1% Ethanol	91
2	20 Mevastatin in 1% Ethanol	88
3	4 Fluvastatin	84
4	20 Lovastatin in 1% Ethanol	83
5	10 Simvastatin in 1% Ethanol	80
6	2 Fluvastatin	79
7	15 Simvastatin in 1.5% Ethanol	79

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
8	5 Simvastatin in 1% Ethanol	74
9	10 Mevastatin in 1% Ethanol	72
10	20 Atorvastatin in 1% Ethanol	71
11	5 Mevastatin in 1% Ethanol	66
12	0.015 Econazole in 1% Methanol	65
13	0.15 Clotrimazole in 1% Methanol	64
14	0.02 Econazole in 1% Methanol	64
15	1 Fluconazole in 0.09 mg/ml NaCl	62
16	0.125 Clotrimazole in 1% Methanol	60
17	0.1 Clotrimazole in 1% Methanol	58
18	2 Fluconazole in 0.09 mg/ml NaCl	52
19	0.03 Econazole in 1% Methanol	51
20	15 Atorvastatin in 1% Ethanol	51
21	3 Fluconazole in 0.09 mg/ml NaCl	50
22	50 Nifedipine in 1% DMSO	39
23	50 Progesterone in 1% DMSO	36
24	10 Progesterone in 1% DMSO	36
25	40 Nifedipine in 1% DMSO	33
26	1.5 Tunicamycin in 1% DMSO	32

Table 8 demonstrates — in accord with results presented in Examples 1 - 4, above — that applying the algorithms of FIGS. 1A, 1B, and 6 to gene expression profiles containing 1532 distinct gene reporters permits quantitation of the relatedness of drugs to 10 µg/ml Lovastatin, an HMG-CoA reductase inhibitor.

Thus, other drugs of the same class — Mevastatin, Fluvastatin, Simvastatin and Atorvastatin —

are shown to be most closely related to Lovastatin. Drugs affecting other steps of the sterol biosynthetic pathway, such as econazole, clotrimazole, and fluconazole, appear next in the ordered list. Drugs with substantially different structure or mode of action, such as progesterone, nifedipine and tunicamycin, follow thereafter. A wide variety of other agents, having even lower relative profile scores, are not shown.

The database of gene expression profiles that was used to generate Table 8 was then queried and subjected to the algorithm schematized in FIGS. 9 and 10. This algorithm is designed to identify a subset of the 1532 genes in the gene expression profiles that is, notwithstanding the reduced number of genes, sufficiently representative of the gene expression repertoire to permit quantitation of the relatedness of the gene expression profiles. In order to achieve a subset with 96 genes — equal to the number of wells of a standard microtiter plate — the range threshold was empirically set to 10 and the correlation threshold to 0.675. The algorithms were implemented on a digital computer, with the algorithmic steps coded in C.

The subset of genes so identified is listed below in Table 9. Functions listed in the table for the genes selected according to the present invention are those functions presently reported in the *Saccharomyces* Genome Database at Stanford University (http://genome-www.stanford.edu/Saccharomyces).

TABLE 9

Gene	Function
PDR12	multidrug resistance transporter; similar to pdr5p
SUC2	invertase
ADH2	alcohol dehydrogenase 2
EUG1	protein disulfide isomerase homolog
YJL105w	
AGA1	anchorage subunit of α-agglutinin
HXT11	glucose permease; high-affinity hexose transporter
YEL065w	
ERG10	acetoacetyl coa thiolase
RPL39	ribosomal protein rpl46 (rat l39)
YGP1	gp37 glycoprotein synthesized in response to nutrient limitation
NUT2	negative regulator of urs two of the ho endonuclease promoter
SNQ2	putative ATP-dependent permease
ECM1	extracellular mutant
YER166w	
MET16	3'phosphoadenylylsulfate reductase
BIO3	7,8-diamino-pelargonic acid aminotransferase
ZEO1	resistance to zeocin
TIF2	translation initiation factor
THI4	thiamine biosynthesis
GLN1	glutamine synthetase
ECM2	extracellular mutant
IDI1	isopentenyl diphosphate:dimethylallyl diphosphate isomerase
PAI3	cytoplasmic inhibitor of proteinase pep4p
ACH1	acetyl coa hydrolase

Gene	Function
YEL047c	
PDR5	multidrug resistance transporter
MFalpha 1	mating factor alpha
CHA1	catabolic serine (threonine) dehydratase
CPA2	carbamyl phosphate synthetase
YER150w	
YJR070c	
HST3	homolog of sir2
GZF3	GATA zinc finger protein 3 homologous to dal80
SPS100	sporulation-specific wall maturation protein
SW14	transcription factor
MFA2	mating α-factor pheromone precursor
SAP155	155 kDa sit4 protein phosphatase-associated protein
TKL2	transketolase, homologous to tkl1
YER073w	
TJL107c	
SED1	putative cell surface glycoprotein
TKL071w	
YBR105c	
FAT2	fatty acid transporter, very similar to fat1
HXT10	high-affinity hexose transporter
CCT7	chaperonin containing t-complex subunit seven
SVS1	vanadate resistance
BUD7	bud site selection
YER064c	

Gene	Function
PIG2	30% identity to protein corresponding to yer054; interacts with gsy2p
YJL181w	
BAR1	a-cell barrier activity on alpha factor
MPT5	
COX6	subunit vi of cytochrome c oxidase
FOX2	peroxisomal multifunctional beta-oxidation protein
GCV2	Glycine decarboxylase complex (P-subunit), glycine synthase (P-subunit), glycine cleavage system (P-subunit)
MIR1	mitochondrial importer receptor (p32); also purified as a mitochondrial phosphate transport protein
YBR147w	
PHO3	acid phosphatase, constitutive
YJL212c	
RPL12A	ribosomal protein rpl15 (yl15) (e. coli l11) (rat l 12b)
YJL017w	
SBA1	Hsp90 (ninety) associated co-chaperone
NIF3	
YHR140w	
YJR105w	
YDR452w	
FET4	Low-affinity fe(ii) transport protein; Putative transmembrane low-affinity fe(ii) transporter
HXT2	high affinity hexose transporter-2
PCL1	G(sub)1 cyclin that associates with pho85
HOM3	aspartate kinase

Gene	Function
TRP2	anthranilate synthase component I
SKI3	Contains 8 copies of the tpr domain; antiviral protein
PHO84	inorganic phosphate transporter, transmembrane protein
PPQ1	protein phosphatase q; may play role in regulation of translation
YER072w	
UTR2	
SBH1	homologous to sbh2p
YER096w	
ILV3	dihydroxyacid dehydratase
YKL078w	
SKT5	protoplast regeneration and killer toxin resistance gene, may be a post-translational regulator of chitin synthase iii activity, interacts with chs3p
YKL187c	
TDH1	glyceraldehyde-3-phosphate dehydrogenase 1
YJR096w	
HIS4	histidine biosynthesis - 3 enzymes
alpha2	in haploid cells, acts with mcm1 to repress a-specific genes. In diploid cells acts with a1 to repress haploid-specific genes.
SER1	phosphoserine transaminase
SIR2	regulator of silent mating loci
OYE3	Nad(p)h dehydrogenase; old yellow enzyme
FIG1	integral membrane protein
TRP1	n-(5'-phosphoribosyl)-anthranilate isomerase
CHS6	Involved in chitin biosynthesis and/or its regulation
CDC8	thymidylate kinase

Gene	Function
MRS6	Rab geranylgeranyl transferase

As can be seen, this subset, selected without regard to gene function, embraces a diverse collection of genes with disparate functions.

The relatedness of each gene expression profile in the database to that produced in the presence of 10 µg/ml Lovastatin was then quantified pairwise, substantially according to the method set forth in FIGS. 1A, 1B and 6, using only the expression data from the 96 genes listed in Table 9. The measures of pairwise relatedness were then ordered, with the following results.

TABLE 10

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
0	10 Lovastatin in 1% Ethanol (index, or reference, condition; experiment 1538)	100
1	5 Lovastatin in 1% Ethanol	92
2	20 Mevastatin in 1% Ethanol	92
3	20 Lovastatin in 1% Ethanol	89
4	10 Simvastatin in 1% Ethanol	84
5	4 Fluvastatin	83
6	2 Fluvastatin	80
7	5 Simvastatin in 1% Ethanol	79
8	10 Mevastatin in 1% Ethanol	79
9	15 Simvastatin in 1.5% Ethanol	79
10	5 Mevastatin in 1% Ethanol	79

Rank	Treatment (drug concentrations in µg/ml)	Relative Profile Score
11	20 Atorvastatin in 1% Ethanol	76
12	15 Atorvastatin in 1% Ethanol	63
13	0.015 Econazole in 1% Methanol	62
14	0.15 clotrimazole in 1% Methanol	61
15	0.125 Clotrimazole in 1% Methanol	59
16	50 Nifedipine in 1% DMSO	58
17	0.02 Econazole in 1% Methanol	58
18	0.03 Econazole in 1% Methanol	55
19	1 Fluconazole in 0.09 mg/ml NaCl	54
20	0.1 Clotrimazole in 1% Methanol	51
21	40 Nifedipine in 1% DMSO	46
22	1 Tunicamycin in 1% DMSO	44
23	1.5 Tunicamycin in 1% DMSO	42
24	2 Tunicamycin in 1% DMSO	41
25	100 Benfluorex hydrochloride in 1% DMSO	40
26	2 Ciclopirox olamine	40

Table 10 confirms that informative subsets of genes may be selected that permit the quantitative analysis of gene expression profiles. As in the analysis presented in Table 8 using data from all 1532 available genes, the analysis presented in Table 10, using only the 96 genes listed in Table 9, identifies HMG-CoA reductase drugs as most closely related to Lovastatin, with drugs acting elsewhere in the same biosynthetic pathway appearing next most closely related, with drugs that are entirely unrelated in target and effect shown as least closely related.

Although this demonstration was performed by selecting 96 genes from among the 1532 genes for which expression data were available in the database, the identification of this informative subset would permit the subsequent, prospective, acquisition of informative gene expression data from only those identified reporters, with confidence that the data so acquired would permit the quantitative analysis of gene expression profiles.

All patents, patent publications, and other published references mentioned herein are hereby incorporated by reference in their entirety as if each had been individually and specifically incorporated by reference herein.

While preferred illustrative embodiments of the present invention are described, it will be apparent to one skilled in the art that various changes and modifications may be made therein without departing from the invention, and it is intended in the appended claims to cover all such changes and modifications which fall within the true spirit and scope of the invention.

What is claimed is:

1. A method of quantifying the relatedness of a first and second gene expression profile, comprising the steps of:

(a) generating a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) formulating a relative expression score for each pair of said first and second gene expression signals; and

(c) calculating from said pair-wise relative expression scores a composite score,

wherein said composite score quantifies the relatedness of the two gene expression profiles.

2. The method of Claim 1, wherein said gene expression signal generating step comprises the steps of:

(a1) comparing the magnitude of an initial expression signal acquired for each of said genes to the magnitude of an initial background signal acquired for its respective gene expression profile; and

(a2) adjusting the magnitude of each of said initial expression signals that is less than its respective initial background signal.

3. The method of Claim 2, wherein said gene expression signal generating step further comprises a subsequent step of:

(a3) normalizing the magnitude of said initial expression signals and said adjusted initial expression signals across all of said signals for the respective gene expression profile.

4. The method of Claim 3, wherein said gene expression signal generating step further comprises a subsequent step of:

(a4) assigning, as the value for each of said gene expression signals, the logarithm of said normalized signal.

5. The method of Claim 4, wherein said gene expression signal generating step further comprises a subsequent step of:

(a5) subtracting, for each of said normalized log signals, an identically processed gene expression signal as acquired for said gene from a condition-matched control.

6. The method of Claim 1 wherein said relative expression score formulating step comprises the steps of:

(b1) calculating, for each pair of said first and second gene expression signals, a ratio therebetween;

(b2) eliminating from further processing each of said calculated ratios for which said earlier steps of background signal adjustment and normalization potentially altered said ratio's direction.

7. The method of Claim 6, wherein said relative expression score formulating step further comprises the subsequent steps of:

(b3) comparing the magnitude of the absolute value of said calculated ratio to the magnitude of a threshold constant; and

(b4) eliminating from further processing each of said calculated ratios the absolute value of which does not exceed said threshold constant.

8. The method of Claim 7, wherein said relative expression score formulating step further comprises a subsequent step of:

(b5) normalizing each of said relative expression scores individually for the maximum expression signal observed historically for said expression score's gene.

9. The method of Claim 6, wherein said relative expression score formulating step further comprises a subsequent step of:

(b3) normalizing each of said relative expression scores individually for the maximum expression signal observed historically for said expression score's gene.

10. The method of any one of Claims 1 - 9 wherein said composite score calculating step comprises the steps of:

(c1) cumulating all of said relative expression scores not previously eliminated; and

(c2) adjusting for the percentage of genes previously eliminated.

11. A method of quantifying the relatedness of a first and second gene expression profile, comprising the steps of:

(a) generating a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) performing a linear regression on the set of paired first and second gene expression signals for said commonly represented genes;

wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

12. The method of Claim 11, wherein said gene expression signal generating step comprises the steps of:

(a1) comparing the magnitude of an initial expression signal acquired for each of said genes to the magnitude of an initial background signal acquired for its respective gene expression profile; and

(a2) adjusting the magnitude of each of said initial expression signals that is less than its respective initial background signal.

13. The method of Claim 12, wherein said gene expression signal generating step further comprises a subsequent step of:

(a3) normalizing the magnitude of said initial expression signals and said adjusted initial expression signals across all of said signals for the respective gene expression profile.

14. The method of Claim 13, wherein said gene expression signal generating step further comprises a subsequent step of:

(a4) assigning, as the value for each of said gene expression signals, the logarithm of said normalized signal.

15. The method of Claim 14, wherein said gene expression signal generating step further comprises a subsequent step of:

(a5) subtracting, for each of said normalized log signals, an identically processed gene expression signal as acquired for said gene from a condition-matched control.

16. The method of Claim 11, wherein said first and second gene expression signals include signals with magnitude less than 2 natural logs.

17. The method of Claim 16, wherein the magnitude of said first and second gene expression signals include signals with magnitude less than 1 natural log.

18. A method of ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising the steps of:

(a) quantifying pairwise the relatedness of each of said plurality of gene expression profiles to said preselected gene expression profile; and

(b) ordering said pairwise-measured quantities.

19. A method of quantifying the relatedness of a first and a second environmental condition upon a cell, comprising the steps of:

(a) obtaining from said cell or from a genotypically identical cell a gene expression profile under each of said first and second environmental conditions; and

(b) quantifying the relatedness of said first and second gene expression profile.

20. The method of Claim 19 wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of any one of Claims 1 - 9.

21. The method of Claim 19 wherein said step of quantifying the relatedness of gene expression profiles

is performed according to the method of any one of Claims 11 - 17.

22. The method of Claim 19 wherein said first and second environmental conditions comprise exposure of said cell to a first and second chemical compound.

23. A method of ordering the relatedness of a plurality of environmental conditions to a single preselected environmental condition upon a cell, comprising the steps of:

(a) obtaining from said cell or from genotypically identical cells a gene expression profile for each of said plurality of environmental conditions and for said preselected environmental condition;

(b) quantifying pairwise the relatedness of each of said plurality of gene expression profiles to said preselected gene expression profile; and

(c) ordering said pairwise-measured quantities.

24. The method of Claim 23, wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of Claim 1.

25. The method of Claim 23, wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of Claim 11.

26. The method of Claim 23, wherein each of said environmental conditions comprises exposure of said cell to a chemical compound.

27. A method of quantifying the relatedness of a preselected environmental condition to a defined genetic mutation of a cell, comprising the steps of:

(a) obtaining a first gene expression profile from a cell bearing said defined mutation and a second gene expression profile from a wild-type cell under said preselected environmental condition; and

(b) quantifying the relatedness of said first and second gene expression profile.

28. The method of Claim 27 wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of Claim 1.

29. The method of Claim 27 wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of Claim 11.

30. The method of Claim 27 wherein said preselected environmental condition comprises exposure of said cell to a chemical compound.

31. A method of ordering the relatedness of each of a plurality of environmental conditions to a defined genetic mutation of a cell, comprising the steps of:

(a) obtaining a set of first gene expression profiles from a wild type cell under each one of said plurality of environmental conditions and a second gene expression profile from a cell having said defined mutation;

(b) quantifying pairwise the relatedness of each of said first gene expression profiles to said second gene expression profile; and

(c) ordering said pairwise-measured quantities.

32. The method of claim 31 wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of claim 1.

33. The method of claim 31 wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of claim 11.

34. The method of claim 31 wherein said environmental conditions include exposure of said cell to a chemical compound.

35. A method of quantifying the relatedness of a first genetic mutation of a cell to a second genetic mutation of a cell, comprising the steps of:

(a) obtaining a first gene expression profile from a cell having said first genetic mutation and a second gene expression profile from a cell having said second genetic mutation; and

(b) quantifying the relatedness of said first and second gene expression profile.

36. The method of claim 35, wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of Claim 1.

37. The method of claim 35, wherein said step of quantifying the relatedness of gene expression profiles is performed according to the method of Claim 11.

38. A method of ordering the relatedness of each of a plurality of genetic mutations to a preselected genetic mutation of a cell, comprising the steps of:

(a) obtaining a set of first gene expression profiles from cells each having one of said plurality of genetic mutations and a second gene expression profile from a cell having said preselected mutation;

(b) quantifying pairwise the relatedness of each of said first gene expression profiles to said second gene expression profile; and

(c) ordering said pairwise-measured quantities.

39. A system for quantifying the relatedness of a first and second gene expression profile, comprising:

(a) means for generating a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) means for formulating a relative expression score for each pair of said first and second gene expression signals; and

(c) means for calculating from said pair-wise relative expression scores a composite score,

wherein said composite score quantifies the relatedness of the two gene expression profiles.

40. A system for quantifying the relatedness of a first and second gene expression profile, comprising:

(a) means for generating a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) means for performing a linear regression on the set of paired first and second gene expression signals for said commonly represented genes;

wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

41. A system for ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising:

(a) means for quantifying pairwise the relatedness of each of said plurality of gene expression profiles to said preselected gene expression profile; and

(b) means for ordering said pairwise-measured quantities.

42. A computer system for quantifying the relatedness of a first and second gene expression profile, comprising a processor programmed to:

(a) generate a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) formulate a relative expression score for each pair of said first and second gene expression signals; and

(c) calculate from said pair-wise relative expression scores a composite score,

wherein said composite score quantifies the relatedness of the two gene expression profiles.

43. A computer system for quantifying the relatedness of a first and second gene expression profile, comprising a processor programmed to:

(a) generate a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) perform a linear regression on the set of paired first and second gene expression signals for said commonly represented genes;

wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

44. A computer system for ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising a processor programmed to:

(a) quantify pairwise the relatedness of each of said plurality of gene expression profiles to said preselected gene expression profile; and

(b) order said pairwise-measured quantities.

45. A computer readable storage medium storing instructions that, when executed by a computer, cause the computer to perform a method of quantifying the relatedness of a first and second gene expression profile, the method comprising:

(a) generating a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) formulating a relative expression score for each pair of said first and second gene expression signals; and

(c) calculating from said pair-wise relative expression scores a composite score,

wherein said composite score quantifies the relatedness of the two gene expression profiles.

46. A computer readable storage medium storing instructions that, when executed by a computer, cause the computer to perform a method of quantifying the relatedness of a first and second gene expression profile, the method comprising:

(a) generating a first and second gene expression signal for each gene commonly represented in said first and second gene expression profiles;

(b) performing a linear regression on the set of paired first and second gene expression signals for said commonly represented genes;

wherein the correlation coefficient of such regression quantifies the relatedness of the two gene expression profiles.

47. A computer readable storage medium storing instructions that, when executed by a computer, cause the computer to perform a method of ordering the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising the steps of:

(a) quantifying pairwise the relatedness of each of said plurality of gene expression profiles to said preselected gene expression profile; and

(b) ordering said pairwise-measured quantities.

48. A computer readable storage medium containing a data structure configured to store data that quantitatively relates a first and second gene expression profile, said data structure comprising an identifier for each of said expression profiles and a scalar, said scalar quantitatively relating said first to said second gene expression profiles.

49. A computer readable storage medium containing a data structure configured to store data that orders the relatedness of a plurality of gene expression profiles to a single preselected gene expression profile, comprising:

(a) an ordered list of scalars, each scalar quantifying pairwise the relatedness of each of said plurality of gene expression profiles to said preselected gene expression profile; and

(b) identifiers that associate each scalar with its respective gene expression profile.

50. A method of selecting an informative subset of genes for expression analysis, comprising:

selecting, from each group of genes whose expression is correlated, the gene with greatest expressive range.

51. The method of claim 50, wherein said selection is made from the set of genes commonly represented in a plurality of gene expression profiles.

52. The method of claim 51, wherein each of said ranges and each of said correlations is calculated from expression data in said plurality of gene expression profiles.

53. The method of claim 52, wherein said range is calculated as the ratio of maximum expression to minimum expression.

54. The method of claim 52, wherein said selecting step comprises the substeps of:

(a) ordering the set of genes commonly represented in said plurality of gene expression profiles from greatest to least in expressive range; and then

(b) choosing the gene with greatest expressive range from each group of genes whose expression in said plurality of gene expression profiles is correlated.

55. The method of claim 53, wherein said choosing substep comprises successive iterations of:

(b1) selecting for said subset the first gene retained in the ordered set that is not yet selected;

(b2) calculating, from said plurality of gene expression profiles, the correlation in expression of each gene in said ordered set to that of the selected gene;

(b3) eliminating from said ordered set all genes with correlation exceeding a threshold value.

56. The method of claim 53, wherein said ordering step further comprises the antecedent step of: eliminating all genes with a range less than a threshold value.

57. A system for selecting an informative subset of genes for expression analysis, comprising:

means for selecting, from each group of genes whose expression is correlated, the gene with greatest expressive range.

58. A computer system for selecting an informative subset of genes for expression analysis, comprising a processor programmed to select, from each group of genes whose expression is correlated, the gene with greatest expressive range.

59. A computer readable storage medium storing instructions that, when executed by a computer, cause the computer to perform a method of selecting an

informative subset of genes for expression analysis, the method comprising selecting, from each group of genes whose expression is correlated, the gene with greatest expressive range.

60. A computer readable storage medium containing a data structure configured to store data that identifies an informative subset of genes for expression analysis, comprising: a set of gene identifiers, optionally including a description of gene function.

61. A method of cellular phenotyping, comprising:
selecting no more than 20% of a cell's expressible genes for expression analysis;
wherein the concurrent expression of said selected genes sufficiently defines said cell's phenotype as to permit said phenotype quantitatively to be related to the phenotype of another cell.

62. The method of claim 61, wherein no more than 10% of said cell's expressible genes are selected.

63. The method of claim 62, wherein no more than 5% of said cell's expressible genes are selected.

64. The method of claim 63, wherein no more than 2% of said cell's expressible genes are selected.

65. The method of claim 64, wherein no more than 1% of said cell's expressible genes are selected.

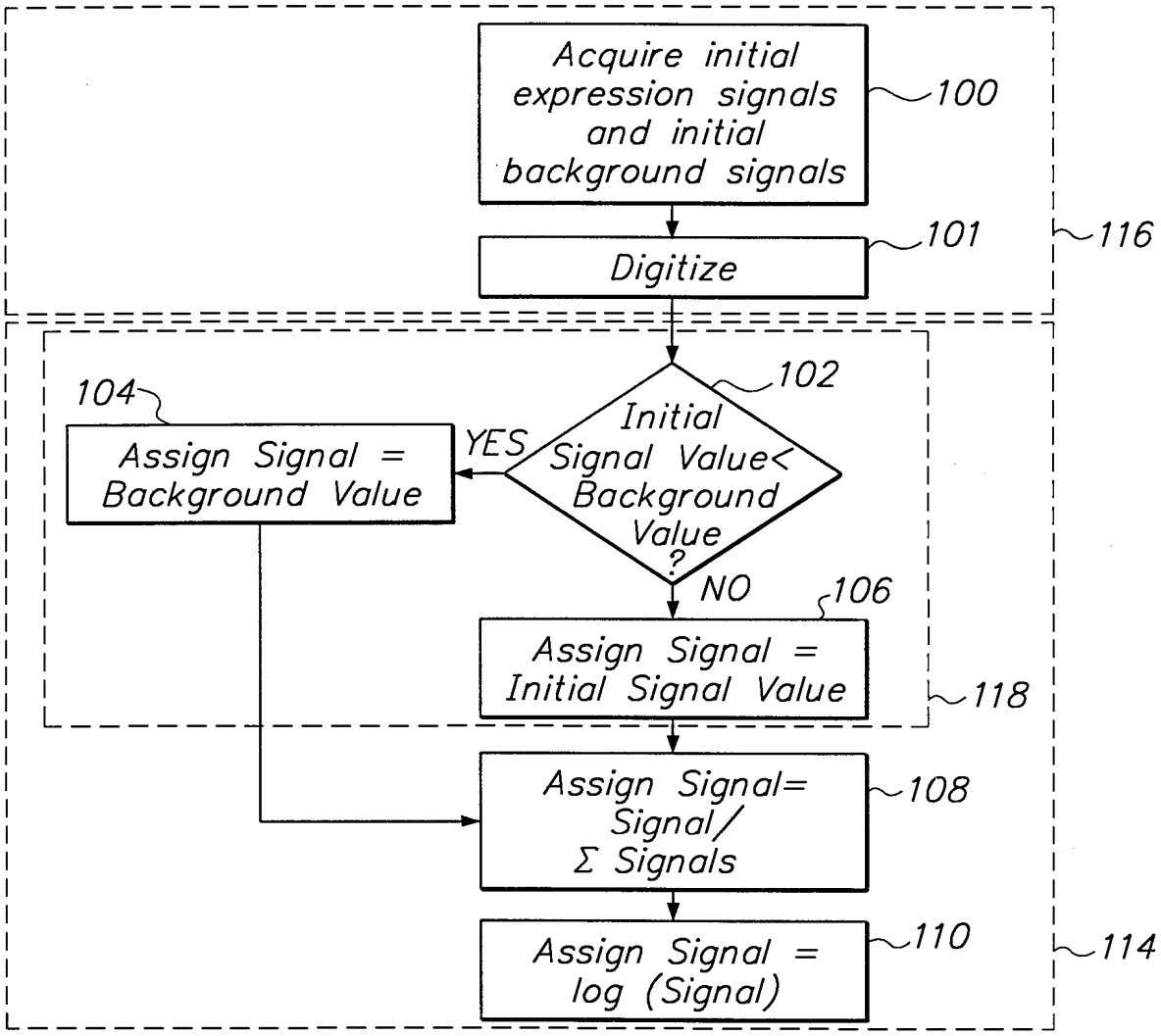
Acquire initial expression signals and initial background signals
100
Digitize
101
116
102
Initial Signal Value< Background Value ?
YES
104
Assign Signal = Background Value
NO
106
Assign Signal = Initial Signal Value
118
Assign Signal= Signal/ Σ Signals
108
Assign Signal = log (Signal)
110
114

FIG. 1A

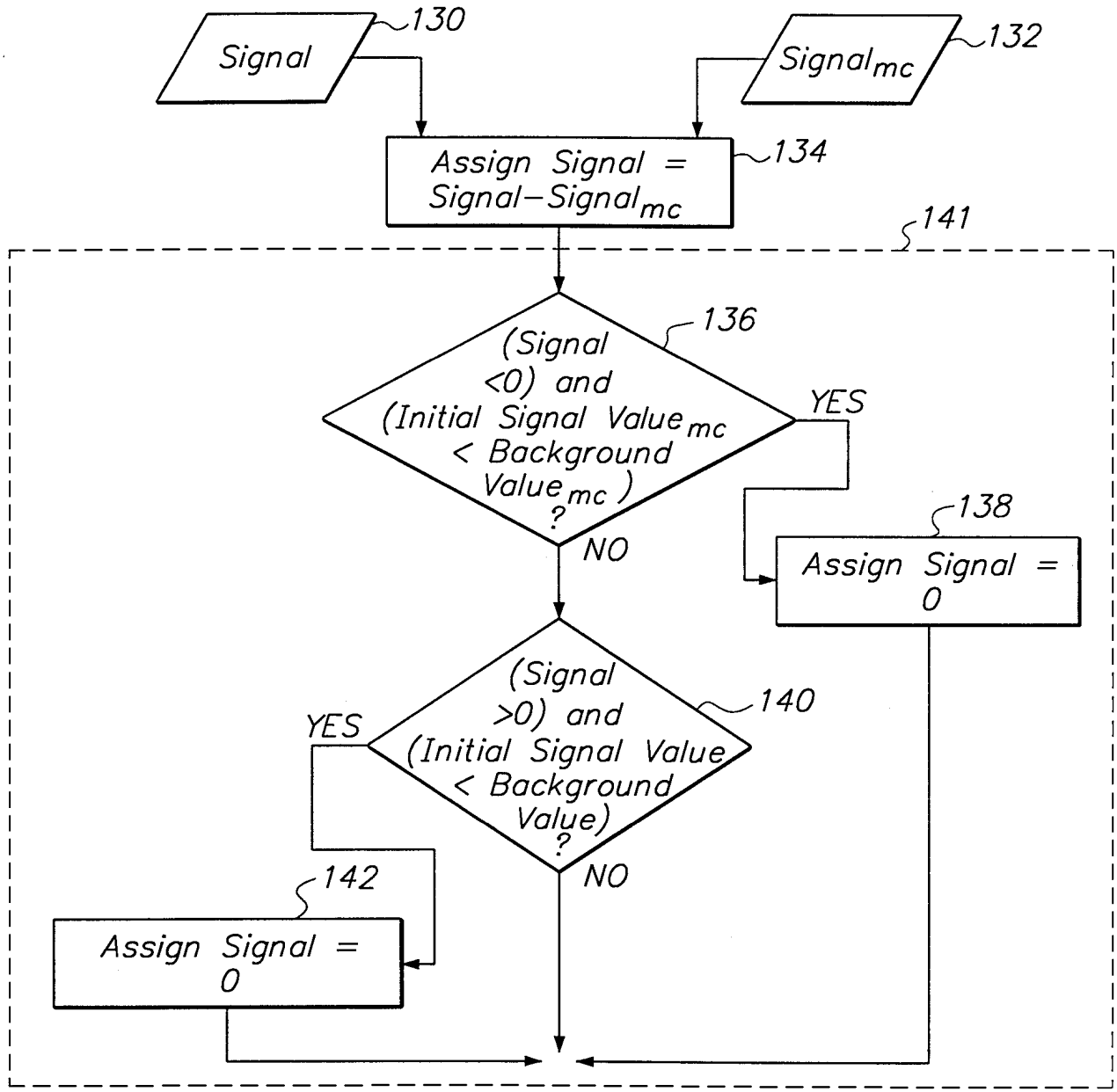
Signal
130
Signal mc
132
Assign Signal = Signal−Signal mc
134
141
(Signal <0) and (Initial Signal Value mc < Background Value mc) ?
136
YES
NO
138
Assign Signal = 0
(Signal >0) and (Initial Signal Value < Background Value) ?
140
YES
NO
142
Assign Signal = 0

FIG. 1B

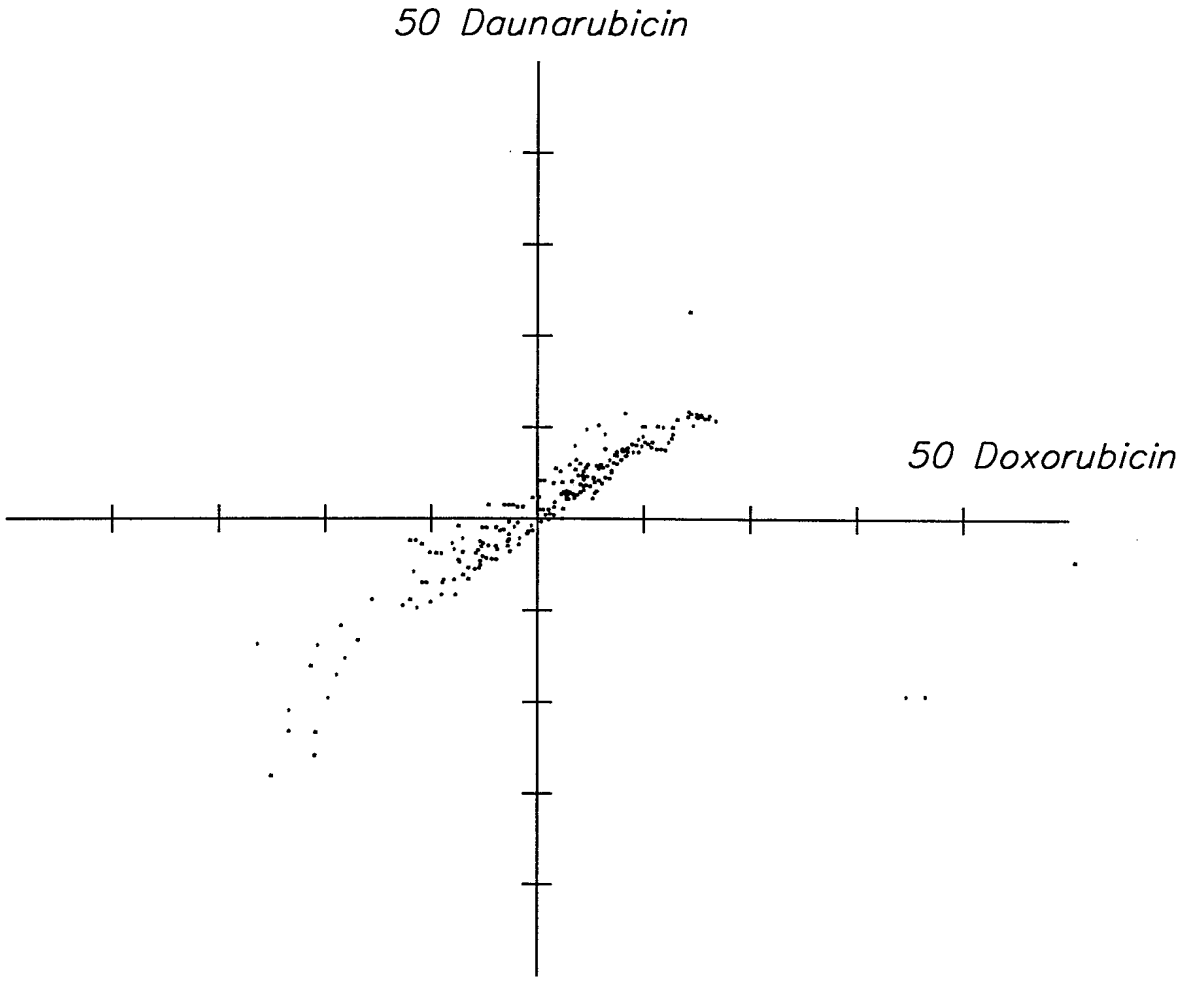

FIG. 2

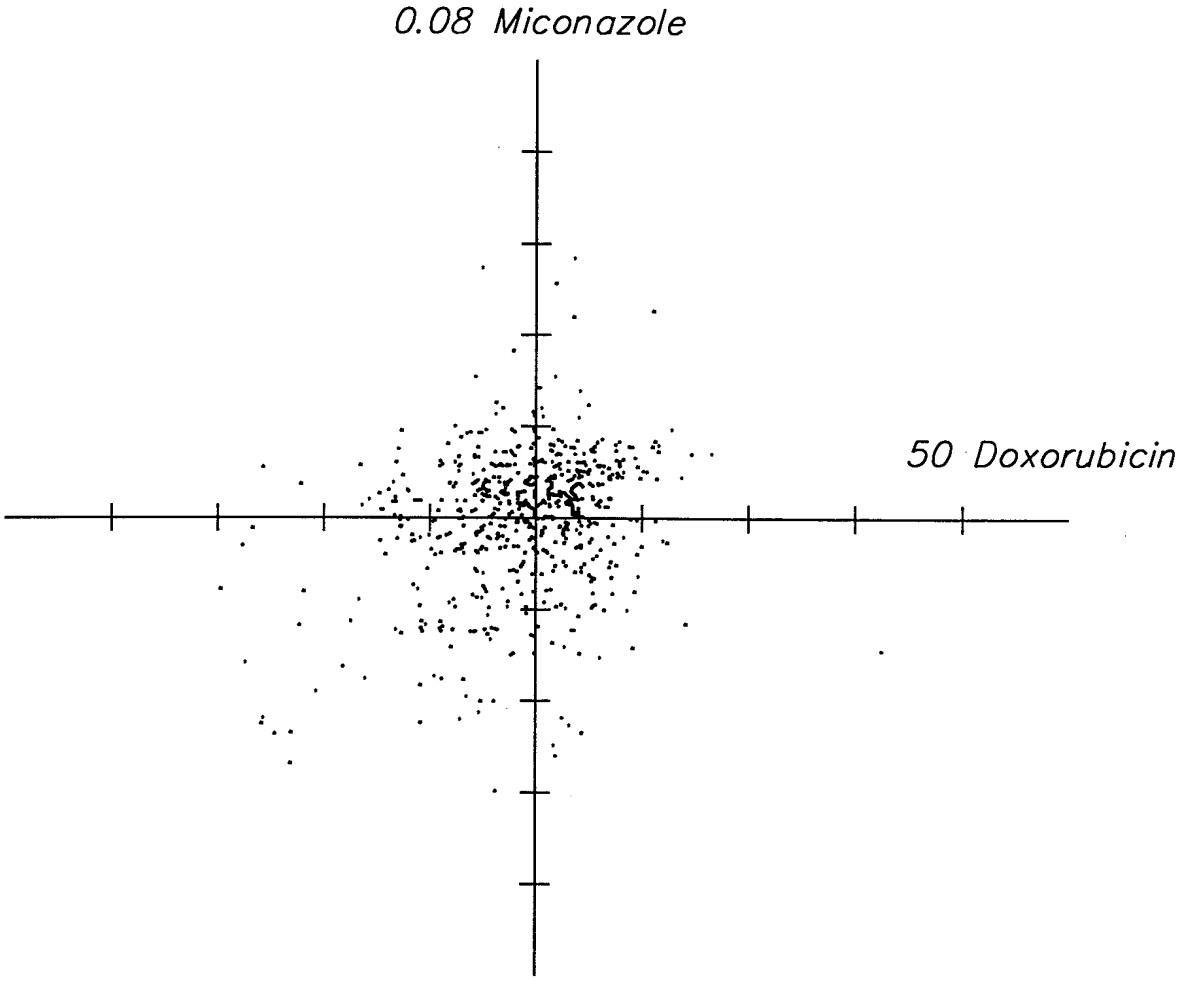

FIG. 3

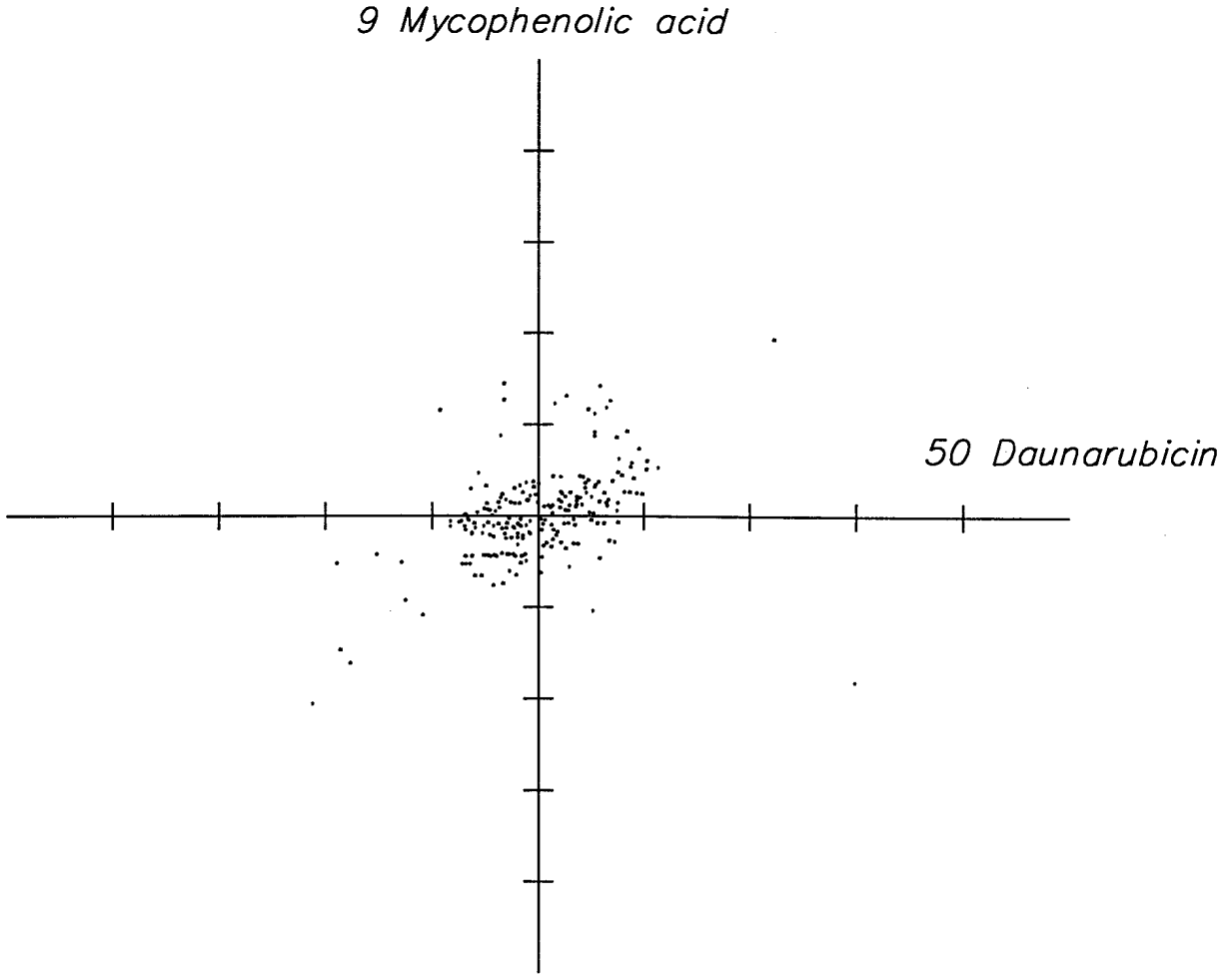

FIG. 4

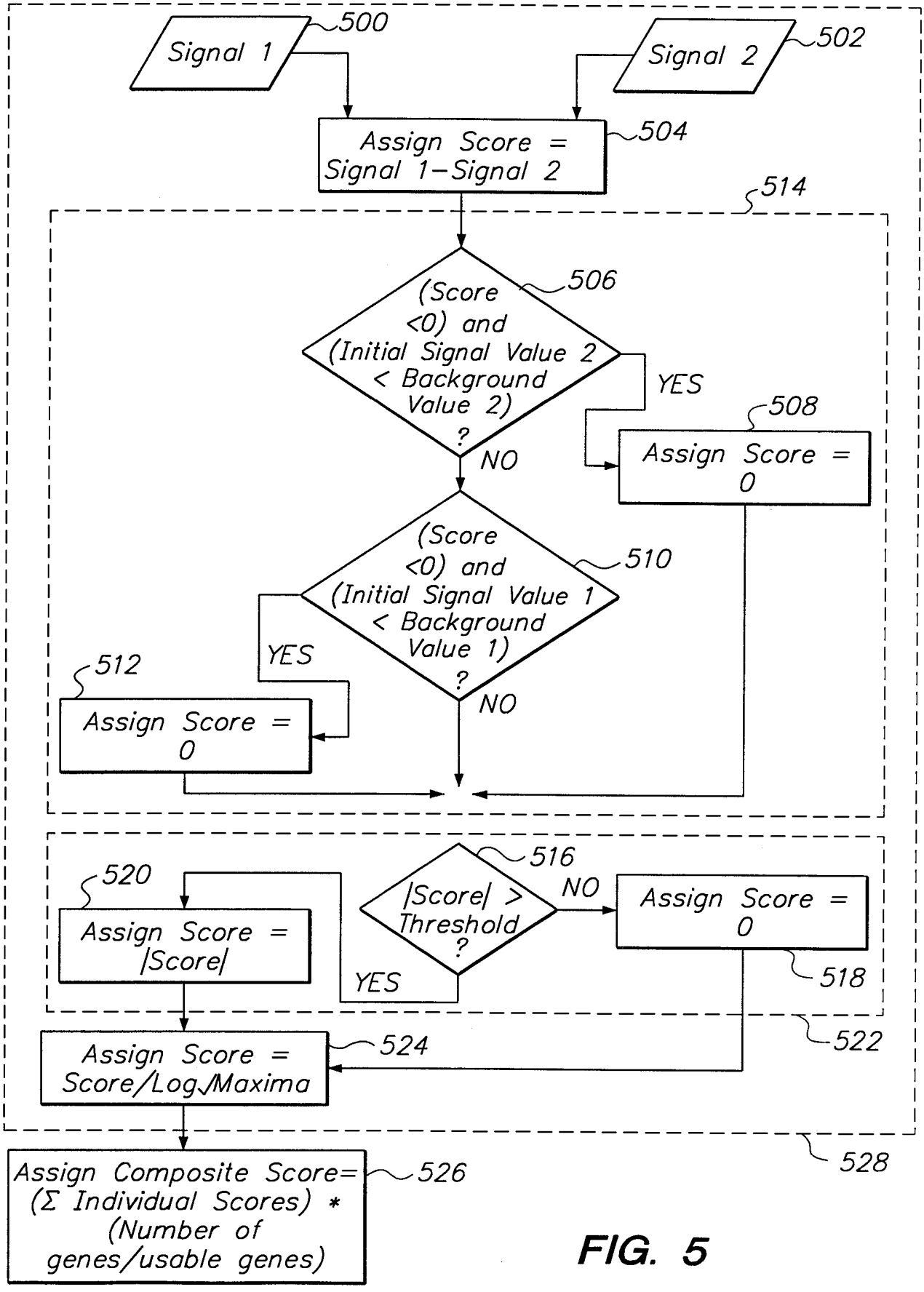
Signal 1
500
Signal 2
502
Assign Score = Signal 1−Signal 2
504
514
(Score <0) and (Initial Signal Value 2 < Background Value 2) ?
506
YES
NO
508
Assign Score = 0
(Score <0) and (Initial Signal Value 1 < Background Value 1) ?
510
YES
NO
512
Assign Score = 0
516
|Score| > Threshold ?
NO
YES
520
Assign Score = |Score|
Assign Score = 0
518
522
Assign Score = Score/Log√Maxima
524
528
Assign Composite Score= (Σ Individual Scores) * (Number of genes/usable genes)
526

FIG. 5

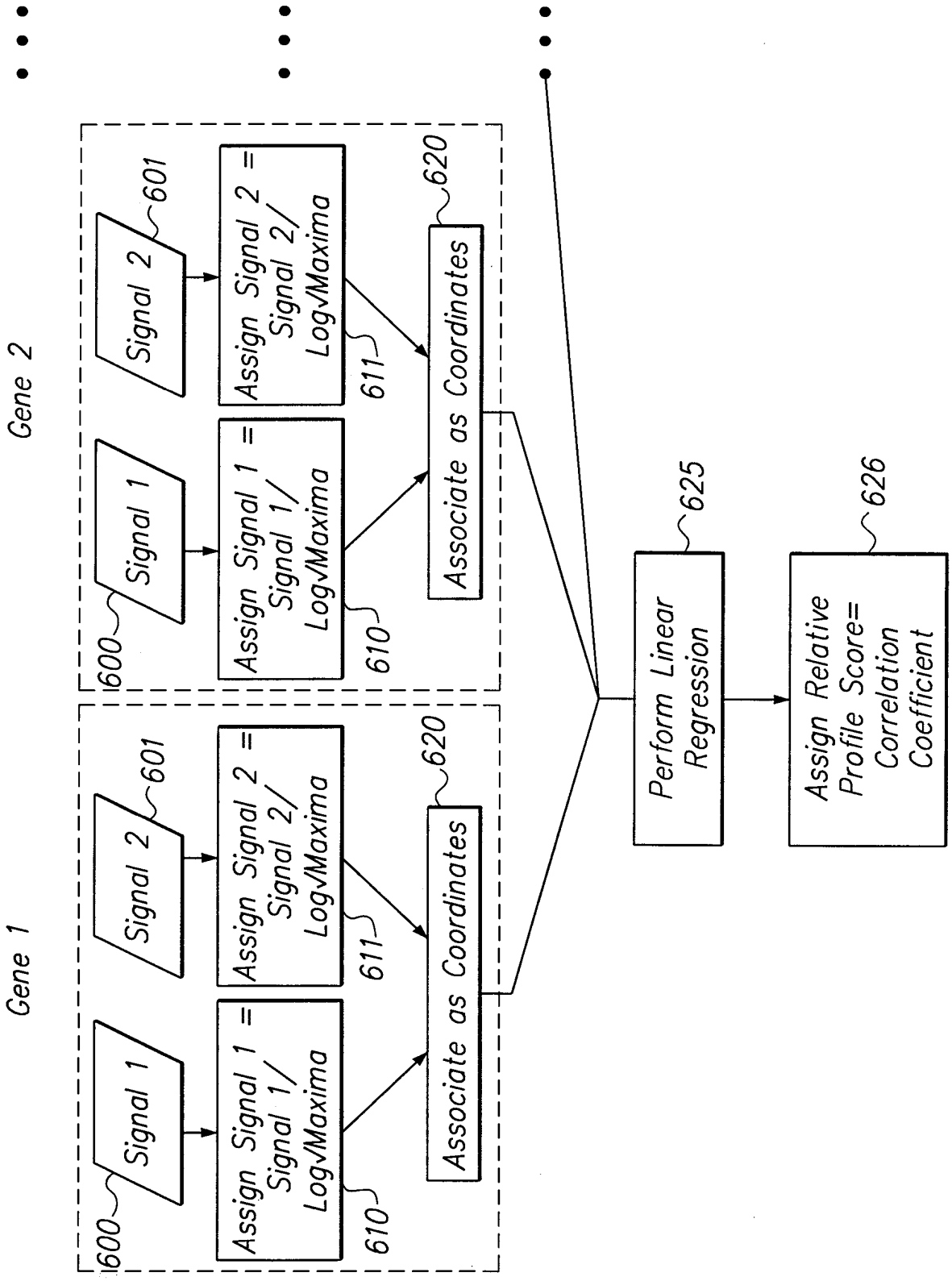
Gene 1
Gene 2
Signal 1
Signal 2
600
601
Assign Signal 1 = Signal 1/ Log√Maxima
Assign Signal 2 = Signal 2/ Log√Maxima
610
611
Associate as Coordinates
620
Perform Linear Regression
625
Assign Relative Profile Score= Correlation Coefficient
626

FIG. 6

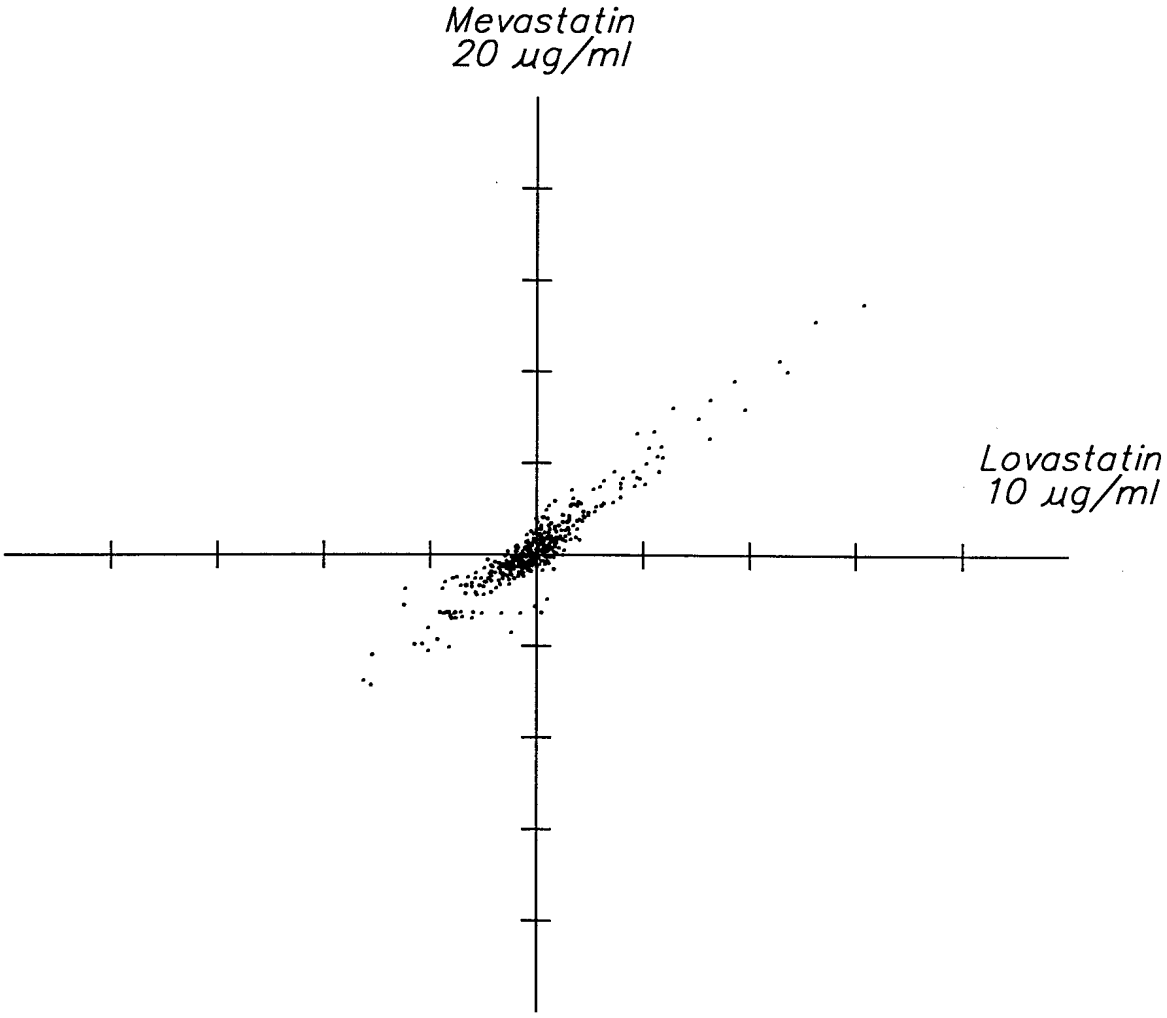

FIG. 7

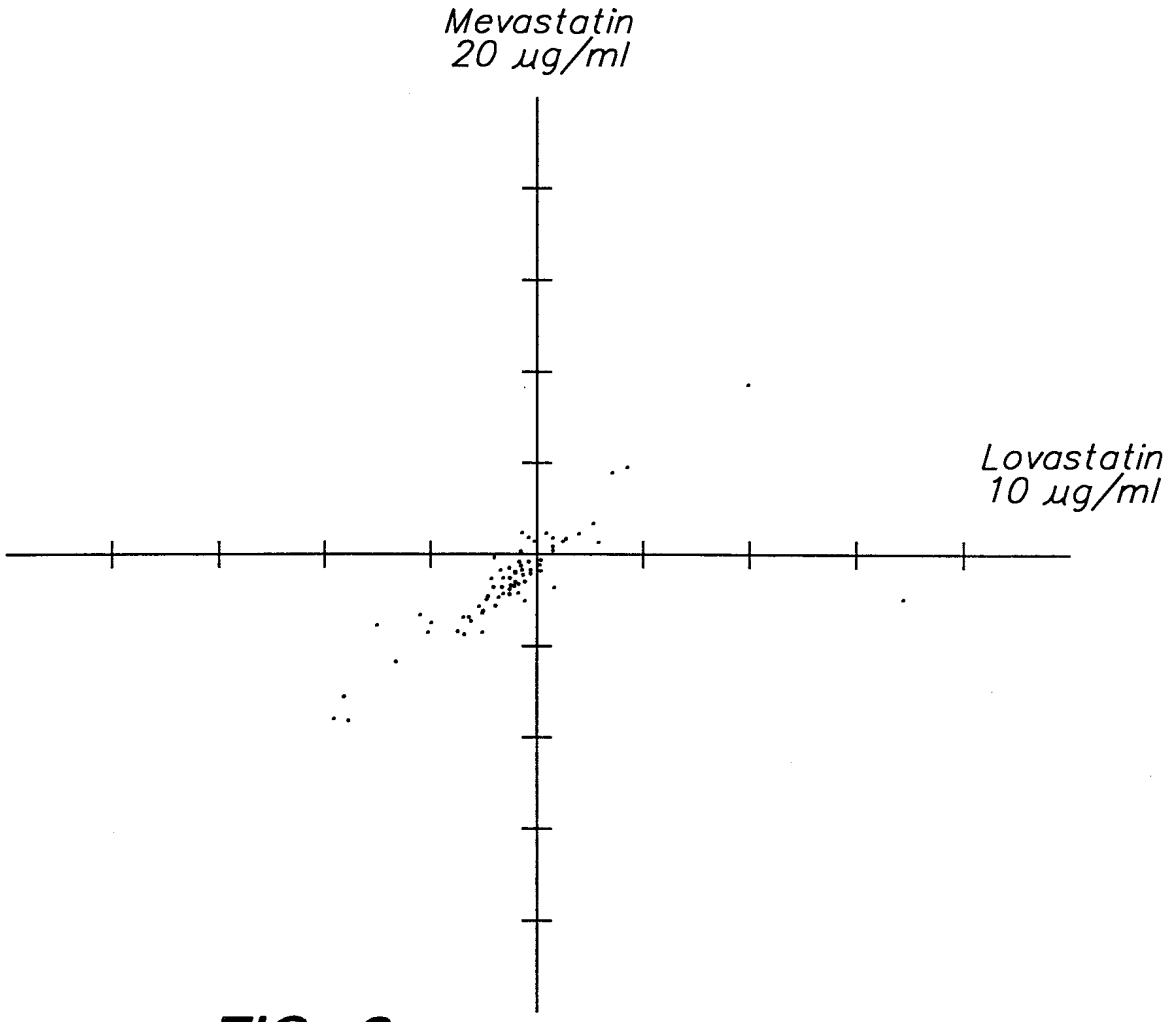
Mevastatin
20 µg/ml
Lovastatin
10 µg/ml

FIG. 8

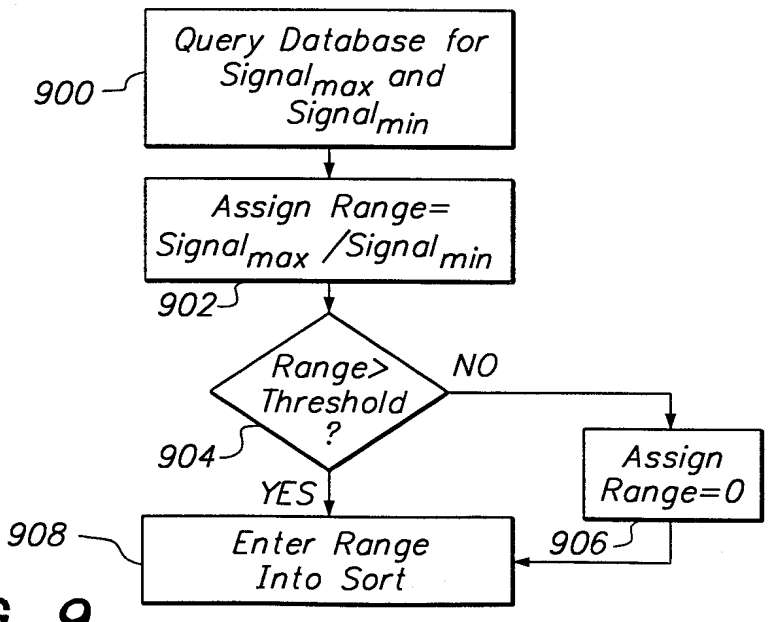
Query Database for Signal$_{max}$ and Signal$_{min}$
900
Assign Range= Signal$_{max}$ /Signal$_{min}$
902
Range> Threshold ?
904
NO
YES
Assign Range=0
906
Enter Range Into Sort
908

FIG. 9

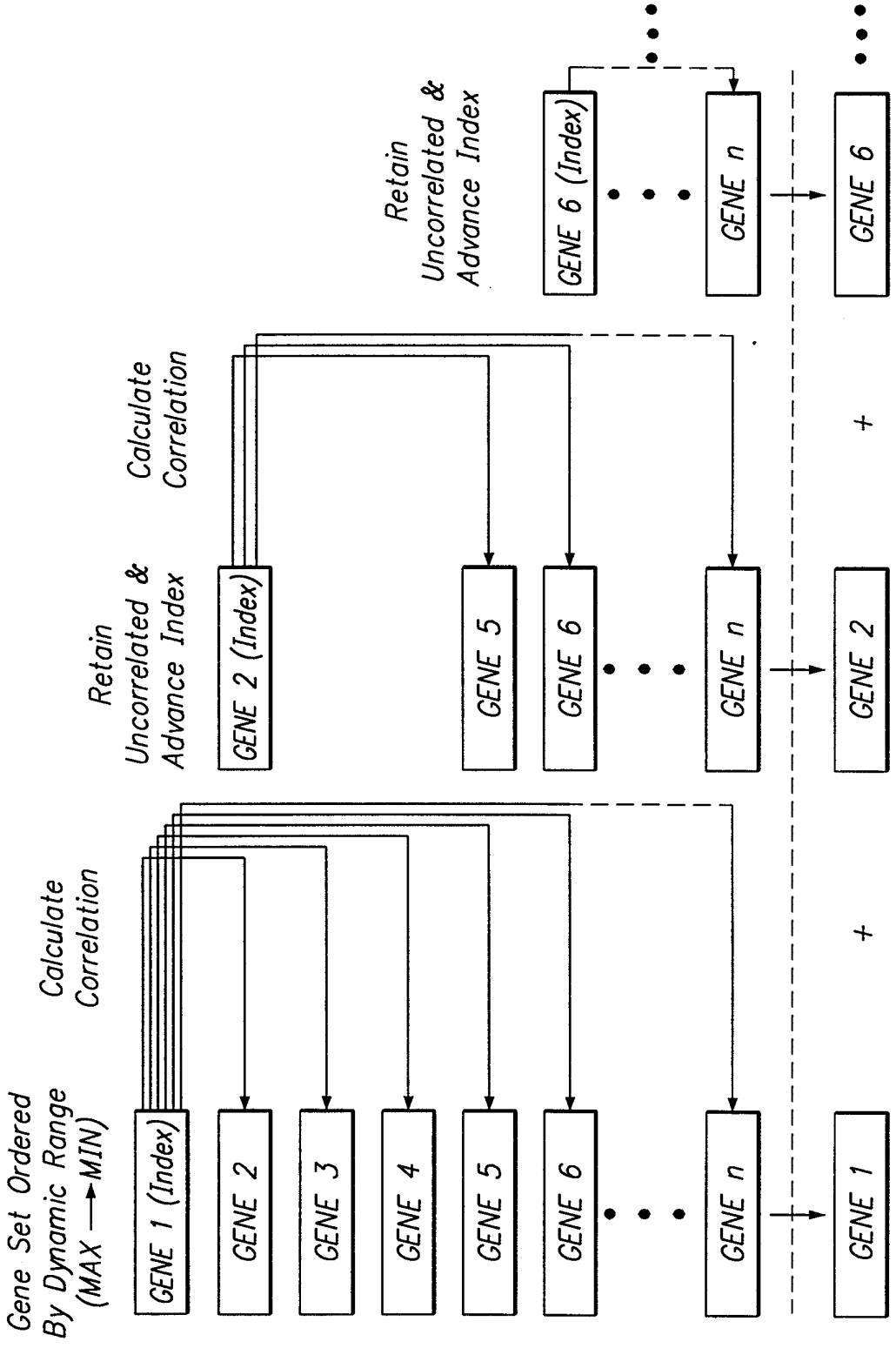
Gene Set Ordered By Dynamic Range (MAX → MIN)
Calculate Correlation
GENE 1 (Index)
GENE 2
GENE 3
GENE 4
GENE 5
GENE 6
GENE n
+
Retain Uncorrelated & Advance Index
GENE 1
Retained Subset
Calculate Correlation
GENE 2 (Index)
GENE 5
GENE 6
GENE n
+
Retain Uncorrelated & Advance Index
GENE 2
Calculate Correlation
GENE 6 (Index)
GENE n
+
GENE 6

FIG. 10

INTERNATIONAL SEARCH REPORT

Interr ıal Application No
PCT/US 99/10387

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68 G06F0/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 06277 A (UNIV CALIFORNIA) 20 February 1997 (1997-02-20) cited in the application see whole doc. esp. claims and examples ---	1-65
X	PIETU G ET AL: "NOVEL GENE TRANSCRIPTS PREFERENTIALLY EXPRESSED IN HUMAN MUSCLES REVEALED BY QUANTITATIVE HYBRIDIZATION OF A HIGH DENSITY CDNA ARRAY" GENOME RESEARCH, vol. 6, no. 6, 1 June 1996 (1996-06-01), pages 492-503, XP000597086 ISSN: 1088-9051 the whole document --- -/--	1-65

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 September 1999	13/09/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Müller, F

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Intern ial Application No
PCT/US 99/10387

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	WO 97 13877 A (LYNX THERAPEUTICS INC ;MARTIN DAVID W (US)) 17 April 1997 (1997-04-17) see whole doc. esp. claims 14ff	1-65
X	WO 97 22720 A (BEATTIE KENNETH LOREN) 26 June 1997 (1997-06-26) see whole doc. esp. claims	1-65
X	WODICKA ET AL: "GENOME-WIDE EXPRESSION MONITORING IN SACCHAROMYCES CEREVISIAE" NATURE BIOTECHNOLOGY, vol. 15, no. 15, December 1997 (1997-12), pages 1359-1367 1367, XP002100297 ISSN: 1087-0156 cited in the application the whole document	1-65
X	WO 98 06874 A (UNIV CALIFORNIA) 19 February 1998 (1998-02-19) cited in the application see whole doc. esp. claims and examples	1-65
A	SCHENA M ET AL: "PARALLEL HUMAN GENOME ANALYSIS: MICROARRAY-BASED EXPRESSION MONITORING OF 1000 GENES" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, vol. 93, no. 20, 1 October 1996 (1996-10-01), pages 10614-10619, XP002912238 ISSN: 0027-8424 the whole document	1-65
A	WO 94 17208 A (XENOMETRIX INC ;FARR SPENCER B (US); MARQUE TODD D (US)) 4 August 1994 (1994-08-04) the whole document	1-65

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Intern ial Application No
PCT/US 99/10387

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9706277 A	20-02-1997	US 5569588 A	29-10-1996
		AU 6720996 A	05-03-1997
		CA 2202154 A	20-02-1997
		EP 0791078 A	27-08-1997
		JP 10507647 T	28-07-1998
WO 9713877 A	17-04-1997	AU 4277896 A	06-05-1996
		AU 6102096 A	30-12-1996
		AU 7717596 A	30-04-1997
		CN 1193357 A	16-09-1998
		CZ 9700866 A	17-09-1997
		CZ 9703926 A	17-06-1998
		EP 0793718 A	10-09-1997
		EP 0832287 A	01-04-1998
		EP 0931165 A	28-07-1999
		FI 971473 A	04-06-1997
		JP 10507357 T	21-07-1998
		NO 971644 A	02-06-1997
		NO 975744 A	05-02-1998
		PL 324000 A	27-04-1998
		WO 9641011 A	19-12-1996
WO 9722720 A	26-06-1997	AU 1687597 A	14-07-1997
WO 9806874 A	19-02-1998	CA 2202152 A	13-02-1998
		US 5777888 A	07-07-1998
		AU 6771596 A	06-03-1998
		EP 0862649 A	09-09-1998
WO 9417208 A	04-08-1994	AT 160178 T	15-11-1997
		AU 692434 B	11-06-1998
		AU 6124394 A	15-08-1994
		DE 69406772 D	18-12-1997
		DE 69406772 T	12-03-1998
		EP 0680517 A	08-11-1995
		ES 2111289 T	01-03-1998
		GR 3026129 T	29-05-1998
		HK 1004920 A	11-12-1998
		JP 9503121 T	31-03-1997
		US 5811231 A	22-09-1998